DOCUMENT RESUME

ED 032 895

LI 001 698

By-Rather, John C., Ed.

Conversion of Retrospective Catalog Records to Machine-Readable Form; A Study of the Feasibility of a National Bibliographic Service.

Library of Congress, Washington, D.C.

Spons Agency-Council on Library Resources, Inc., Washington, D.C.

Pub Date 69

Note-242p.

Available from-Superintendent of Documents, U.S. Government Printing Office, Washington, D.C. (LC 1.2: M 18/5, $2.25).

EDRS Price MF-$1.00 HC Not Available from EDRS.

Descriptors-Administrative Problems, *Bibliographic Citations, *Cataloging, Computer Programs, Computers, Costs, *Data Bases, Feasibility Studies, *Information Processing, *Input Output, National Programs

Identifiers-Library of Congress, *Machine Readable Cataloging, MARC

This report examines (1) the present state of the art of hardware and software applicable to large-scale conversion, storage, and retrieval of retrospective bibliographic information; (2) the organizational and administrative aspects of the task; (3) costs of hardware, software, and manpower; and (4) possible approaches to the timing and funding of the project. The main body of the report examines the various problems involved and explores possible solutions. It is concluded that (1) the MARC (Machine Readable Cataloging) service should be expanded to cover all languages and forms of material; (2) conversion for a national bibliographic data base requires standardization of bibliographic content and machine format; and (3) large-scale conversion should be accomplished as a centralized project. Appendices discuss (1) duplication (overlap) in U.S. library collections, (2) actual and planned data conversion activities in selected libraries, (3) a summary of interviews with consultants, (4) the history and future of Library of Congress (LC) catalog records, (5) changes in LC catalog cards, (6) completeness of machine-readable catalog records, (7) format recognition, (8) computer requirements for a national bibliographic service, and (9) staff complement and unit costs. (JM)

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

ED0 32895

THE LIBRARY OF CONGRESS

CONVERSION OF RETROSPECTIVE CATALOG RECORDS TO MACHINE-READABLE FORM

LI 001698

ERRATA

The calculation of the estimated costs of the computer configuration for second and third shifts (pages 69-72) was based on erroneous information. More accurate budgetary estimates for the basic configuration (p. 69) are: two shifts, $49,000 per month; three shifts, $53,000. For the additional storage (p. 70), the two-shift estimate should be $16,000 and the three-shift estimate, $17,000. As a result, the following corrections should be made:

Page	Paragraph	Line	Correction
71	2	5	$53,000
72	1	3	$53,000
		4	17,000
		5	104,000
	3	1	$49,000
		2	16,000
		3	97,000
	4	2	$7.0 million
		4	3.8 million
100	3	6	$49,000
		9	81,000

The error is regretted.

ERIC/CLIS
OCT 6 '69
UNIVERSITY OF MINNESOTA

CONVERSION OF RETROSPECTIVE CATALOG RECORDS TO MACHINE-READABLE FORM

A Study of the Feasibility of a National Bibliographic Service

Prepared by the RECON Working Task Force

Henriette D. Avram, Chairman

William R. Nugent, Josephine S. Pulsifer, John C. Rather

Joseph A. Rosenthal, Allen B. Veaner

LIBRARY OF CONGRESS · WASHINGTON · 1969

Edited by John C. Rather

L.C. Card 70-601790

For sale by the Superintendent of Documents, U.S. Government Printing Office
Washington, D.C. 20402 - Price $2.25

TABLE OF CONTENTS

Foreword v

Acknowledgments viii

1 INTRODUCTION 1

2 MAJOR CONCLUSIONS AND RECOMMENDATIONS 10

3 USES OF CONVERTED BIBLIOGRAPHIC DATA 13

4 MASTER DATA BASE 20

5 TECHNICAL ALTERNATIVES: MACHINE CONSIDERATIONS 39

6 TECHNICAL ALTERNATIVES: MANPOWER CONSIDERATIONS 74

7 COSTS OF CONVERSION 97

8 FUNDING AND OTHER SUPPORT CONSIDERATIONS 102

APPENDIXES

A Duplication in U. S. Library Collections 106

B Actual and Planned Data Conversion Activities in Selected Libraries and Their Use of Library of Congress Cataloging 111

C Summary of Interviews with Consultants 125

D Library of Congress Catalog Records: Past and Future 136

E Changes in Library of Congress Catalog Cards: Their Extent, Method, and Types 141

F Completeness of Machine-Readable Catalog Records 163

G Format Recognition 169

H Computer Requirements for a National Bibliographic Service 183

I Staff Complements and Unit Costs 224

Index 227

FOREWORD

As a result of the MARC Pilot Project, growing acceptance of the MARC II format, and the implementation of the MARC Distribution Service, libraries throughout the country are beginning to discuss and, in some instances, to plan the conversion of their catalog records to machine-readable form. Since funds and manpower available for this purpose vary among libraries and their bibliographic needs are not always similar, the machine-readable products of uncoordinated conversion projects would differ with respect to completeness and uniformity. Local conversion would also result in a great deal of duplication of bibliographic information about the same items. Not only do the consequences of these eventualities appear economically unsound but also they have serious implications for future plans to create a national data base of bibliographic information in machine-readable form.

The Library of Congress has accepted the responsibility for the conversion of its current cataloging to the MARC II format. The Library is also conducting studies to determine the feasibility of converting its retrospective material. In view of widespread interest, it seemed timely and appropriate to take a closer look at the problems of centralized conversion of retrospective cataloging records and their distribution to the entire library community from a central source. If a workable plan could be conceived and implemented, the machine-readable records would be consistent, the cost savings would be significant, and the first steps toward creation of a national data base would have been taken.

When the Library of Congress presented a proposal for a study of this problem to the Council on Library Resources, Inc., the Council was quick to recognize the far-reaching significance of the undertaking by providing funds without delay. An advisory committee composed of members of the library profession was appointed to provide guidance for the study which was dubbed RECON (REtrospective CONversion). Direct responsibility for the study was assigned to a working task force composed of librarians and systems analysts representing different types of libraries. Henriette D. Avram was chosen to chair the working task force because she conceived the idea for the study and wrote the proposal for the Library of Congress.

Despite the many days devoted to the study, all of the members served on the RECON Working Task Force without compensation and their parent organizations generously allowed time for this purpose. This willingness to contribute the service of experienced personnel does great credit to everyone concerned. It enabled the task force to take a long, hard look at the manifold problems of large-scale conversion of retrospective cataloging records. It is hoped that the findings will benefit the library community and lay the foundation for further planning in this area.

John G. Lorenz
Deputy Librarian of Congress
Officer-in-Charge, RECON Study

RECON STUDY

Officer-in-Charge: John G. Lorenz, Deputy Librarian
of Congress

Working Task Force:

Mrs. Henriette D. Avram, Chairman
Library of Congress

William R. Nugent
Inforonics, Inc.

Mrs. Josephine Pulsifer
Washington State Library

John C. Rather
Library of Congress

Joseph A. Rosenthal
New York Public Library

Allen B. Veaner
Stanford University Libraries

Advisory Committee:

John G. Lorenz, Chairman

Scott Adams
Deputy Director
National Library of Medicine

Col. Andrew A. Aines, Chairman
Committee on Scientific and
Technical Information
Federal Council for Science
and Technology

Herman H. Fussler
Director
University of Chicago

James W. Henderson
Chief, Reference Department
New York Public Library

Frederick G. Kilgour
Director
The Ohio College Library Center

Abraham Lebowitz
Assistant to the Director
National Agricultural Library

Maryan E. Reynolds
State Librarian
Washington State Library

Rutherford D. Rogers
Director of Libraries
Stanford University

Russell Shank
Director of Libraries
Smithsonian Institution

James E. Skipper
University Librarian
University of California, Berkeley

ACKNOWLEDGMENTS

The RECON study provides a notable example of cooperative effort to explore a critical library problem. The Council on Library Resources, Inc., and its president, Fred C. Cole, responded with alacrity to a request for funds to support the study. The time of the members of the working task force was donated by each member's organization. Thanks for this generosity are due L. Quincy Mumford, Librarian of Congress; Lawrence F. Buckland, President, Inforonics, Inc.; Edward G. Freehafer, Director, The New York Public Library; Maryan E. Reynolds, State Librarian of Washington; and Rutherford D. Rogers, Director, Stanford University Libraries.

The members of the advisory committee took time from their busy schedules to attend two day-long meetings and to comment on the development of the projects. The deliberations of the advisory committee were enhanced by the participation of the following representatives of major funding agencies: Burton W. Adkinson, Head, Office of Science Information Service, National Science Foundation; Lee G. Burchinal, Director of Information Technology and Dissemination, Bureau of Research, U. S. Office of Education; Dr. Cole; and Foster E. Mohrhardt, Program Officer, Council on Library Resources, Inc.

Experts in the field of library automation responded generously

to the RECON Working Task Force's solicitation of their opinions on the conversion problem. All of them, gave their advice and comments freely. The list of their names appears with the summary of their views in Appendix C.

Seventy libraries agreed to be interviewed by representatives of Herner and Company on behalf of the RECON Working Task Force. The willingness of these libraries to share their experiences and plans concerning conversion of catalog records to machine-readable form provided significant insights into the problem. The names of the libraries appear in Appendix B.

Although it is not feasible to name every Library of Congress staff member who gave assistance or was consulted during the investigation, the working task force is especially indebted to Susan C. Biebel. Her skill in developing statistical data and her diligence in carrying out a multitude of assignments made an essential contribution to the study and the final report.

In addition to the library survey, two other contractual efforts were used in the RECON study. Coyle and Stewart prepared an analysis of computer requirements that provided a starting point for Appendix H. Use was made also of a study of optical character recognition and conversion devices and procedures by Auerbach and Company which was done concurrently at the Library of Congress.

Special thanks are due John A. Bayless of Planning Research Corporation who donated a day of his time to share his experience in large-scale file conversion.

The evidence of genuine interest in the problem of retrospective conversion was apparent everywhere and there was no lack of opinion about how it might be resolved. The working task force is grateful to the many individuals and organizations who contributed to the study in one way or another. It is to be understood, of course, that the working task force assumes full responsibility for the use made of this information.

Chapter 1

INTRODUCTION

As libraries develop their plans for automation, it becomes increasingly apparent that the full benefits of the computer cannot be realized unless large stores of bibliographic information are available in machine-readable form. The MARC Distribution Service inaugurated by the Library of Congress will provide a source of current cataloging data that, as time, resources, and technology permit, can be expanded to cover virtually all of the Library's current output. Although this may take care of the future, the task of converting the large masses of cataloging information produced during the last 70 years still must be faced.

To accomplish both types of conversion, several critical problems must be solved:

1. Identification of user needs for retrospective cataloging data.

It is obvious that libraries cannot base their products and services solely upon records to be created from this day forward; the bibliographic responsibilities of libraries extend into the past as well as the future. Is a retrospective machine-readable data base needed to service these responsibilities? If so, how shall it be obtained? What might it cost? How would it be distributed? These are some of the questions which immediately arise.

2. The means of maintaining standardization of the machine format for machine-readable catalog records so that libraries can exchange information in this form.

Even in manual systems based upon card and book catalogs, the effective interchange and communication of bibliographic data depends on standardization. Owing to the computer's intolerance of ambiguity in source data, the future transmission and exchange of machine-readable records will be even more dependent upon standardization. Acceptance of MARC II as a standard communication format will provide a common currency for machine-readable catalog records that will perform much the same function as Library of Congress printed cards have done for over two-thirds of a century.

3. The technical requirements for large-scale storage and retrieval of the data store.

Bibliographic data by its nature presents problems in machine input, processing, and output that differ markedly from those posed by numeric data or even by straightforward alphabetic text. The development of the MARC system and the important work at libraries elsewhere have contributed greatly to the solution of these problems. Nevertheless, the requirements for large-scale conversion operations demand further study and, in some cases, implementation must await successful development of new equipment.

4. The systems design and the necessary software required to create, maintain, and disseminate information from a large data base.

Much has been said and written about network concepts and national data bases, but the discussions have been largely at a level divorced from specifics. A pioneering effort is required to plan and implement an actual system. The dynamic nature of bibliographic files creates updating problems of great magnitude. In general, bibliographic records do not become archival; they must be capable of being accessed regardless of their age. To achieve maximum flexibility in retrieving information from a large data base of bibliographical records, it is necessary to provide more than one form of access to the information. These and other problems require the design of file organization and searching techniques that will allow for the most efficient retrieval of records from a large data base. The planning and design of distribution services also requires a major programming effort to handle the many logistic problems.

5. The staffing and funding requirements for a major conversion project.

Capturing retrospective bibliographic information in machine-readable form--to the point where a significantly useful data store will be available--is not a matter of arriving at standards, determining priorities, and developing hardware and software techniques alone. The administrative and personnel framework must be designed and the means of financing all aspects of the task envisioned before such a project can be contemplated as a part of ongoing library operations, whether undertaken at one or many institutions.

Although conversion of retrospective records has been discussed in various contexts[1], these problems have never been fully explored. In view of their magnitude, it seems intuitively clear that a centralized effort to create a data base of retrospective catalog records for national use would have significant benefits in terms of the time, effort, and money to achieve the desired result. On the other hand, decentralized efforts would carry heavy penalties.

Since funds and manpower for automation vary widely among libraries and their needs for bibliographic description are not always similar, the machine-readable records resulting from individual projects will probably reflect varying degrees of completeness. The economic penalties associated with nonstandardized bibliographic procedures are familiar to library administrators. The purported need to deviate from standards in favor of local practices could readily be defended as long as little was known of the costs associated with creation of a custom-tailored bibliographic product. But management's relentless attention to cost-effectiveness is gradually exposing hidden costs and the built-in

---

1. See, for example, De Gennaro, Richard. A strategy for the conversion of research library catalogs to machine-readable form. College and research libraries, v. 28, July 1967, 253-257.

record keeping and accounting functions of computer services provide evidence of inefficiencies.

A principal component cost of any computer system is software development. Indeed, software development and maintenance investments frequently surpass the costs of machine-processing time. This suggests that, in the absence of a national program for conversion, many libraries might undertake to develop essentially the same software at great individual cost. There is, moreover, the danger that independent efforts would result in incompatible record formats and variations in the content of the records that would inhibit effective, economical utilization of networks for the future communication of bibliographic data. Therefore, the question naturally arises: Why not write the software once, convert a full, accurate, up-to-date record, and distribute a standardized product, all on a centralized basis?

The National Program for Acquisitions and Cataloging has provided within a period of only a few years (since 1966) a significant increase in the amount of cataloging data for foreign language titles available from a central source. This suggests that a similar centralized approach for retrospective data through the Library of Congress might satisfy the library community's need in this area. If this conversion effort could comprehend the needs of other libraries as well as those of the Library of Congress, it should result in a true national data base characterized by accuracy, consistency, and economy of production.

The present study undertakes to examine in detail:

1. The present state of the art of hardware and software applicable to large-scale conversion, storage, and retrieval of retrospective bibliographic information.
2. The organizational and administrative aspects of the task, including considerations of which existing files are most suitable for conversion, which segments of those files should have priority for conversion, and how best to accomplish the job.
3. Costs of hardware, software, and manpower for such a project.
4. Possible approaches to the timing and funding of the project; and areas that need intensive additional study.

The complexity of the concept of conversion of retrospective catalog records has affected both the organization and the substance of this report. The main body of the report examines the various problems involved, explores possible solutions, and offers recommendations for action. Supporting studies and documentation are given in the appendixes. These include: (1) reports of consultations with knowledgeable and interested individuals and organizations other than the working task force and the advisory committee; (2) statistical reports substantiating certain conclusions embodied in the report (e.g., duplication of library collections, changes in Library of Congress cards); (3) extended descriptions of fundamental concepts (e.g., completeness of machine-readable catalog records, format recognition), which are only summarized in the report itself; and (4) detailed presentations (e.g., unit costs, machine configurations) elaborating certain aspects of the proposals developed in the course of the study.

The exceedingly wide range of possible alternatives at almost every step of this study forced the working task force to make certain choices and assumptions that deserve to be stated for the reader. The technologies discussed are either operative or in the process of actual development. Proposals for the organization, design, and goals of a conversion project are made within the framework of an attainable system that would result in a product of general utility. Nevertheless, this report does not pretend to be a definitive blueprint of a fully conceived conversion project. Both the brief span of the study and the many uncertainties about specific details made it impossible to do more than provide a broad outline of the problems and how they might be solved. It is hoped, that the report provides a solid foundation for further development and implementation of a workable project.

This study has focused on the feasibility of the conversion of catalog records to machine-readable form as a centralized effort by analyzing some of the problems that must be solved. It has not attempted to predict all of the ways that these records would or could be used once they have been created, although a general discussion of some possible uses of machine-readable records is given in chapter 3. The question of the utility of machine-readable records is relevant not only to retrospective records: it applies equally to current records that are being converted. Therefore, although the question should be studied, it was considered to be out of the scope of the present investigation.

There are, in fact, many problems that are common to all machine-readable records whether current or retrospective. Cataloging rules, provision for filing arrangement, representation of nonroman or other special characters, and techniques for organizing and using large machine files raise important questions that merit study. All of these problems are being or should be investigated but they were considered only tangentially in this report because of the primary emphasis on the problems of converting existing catalog records as they now stand.

In addition, it was considered beyond the scope of the present study to investigate all of the problems inherent in the maintenance and use of a national bibliographic system. The full realization of the benefits of such a system will depend on the accumulation of practical experience in the organization, maintenance, and use of large bibliographic data files and intensive effort in system design.

This study shows that there is widespread interest in conversion, an appreciable amount of ongoing activity (in both actual conversion and in the development of techniques directly applicable to the task), and evidence that many libraries would be willing to follow common standards (such as the MARC II format and uniform cataloging practices). To insure the success of a conversion effort, there must be not only general acceptance of these and other standards, but also a willingness on the part of libraries and the professional associations in the field to give a high priority to the search for funds adequate to insure a product of value in the foreseeable future. It is vital to realize that any coordinated effort to convert retrospective bibliographical information must elicit strong support from the library community.

Chapter 2

MAJOR CONCLUSIONS AND RECOMMENDATIONS

A. General Conclusions

1. The MARC Distribution Service should be expanded to cover all languages and all forms of material as rapidly as resources and technology allow. There should be no conversion of any category of retrospective records until that category is being currently converted.

2. Conversion of some portion of retrospective records to machine-readable form should be an early goal of library automation efforts.

3. Conversion for a national bibliographic data base requires standardization of bibliographic content and machine format. Standards for conversion of retrospective records should be the same as those for current records.

4. The highest priority for retrospective conversion should be given to records most likely to be useful to the largest number of libraries. As nearly as possible, subsequent priorities should be determined by the same criteria.

5. Large-scale conversion should be accomplished as a centralized project. Decentralized conversion would be more costly and unlikely to satisfy requirements for standardization. The project should be under

the direction of the Library of Congress.

B. Specific Recommendations

1. The records to be converted should in effect be those in the LC Official Catalog. Actual conversion would require a two-step process: conversion of portions of the Card Division record set followed by updating the records from the Official Catalog.

2. The initial conversion effort should be limited to English language monograph records issued from 1960 to date. Second priority should be given to Romance and German language monograph records issued from 1960. Third priority should be given to English language monograph records issued from 1898-1959.

3. To meet the emerging needs of libraries, every effort should be made to convert priority one and two records within four years. Conversion of these and other records should start with the most recent year and proceed backward in reverse chronological order.

4. Initially, the method of conversion should involve:
    a. Partial editing of entries from the record set prior to input.
    b. Conversion by magnetic tape inscriber.
    c. Application of a format recognition program.
    d. Comparison of records with the LC Official Catalog.
    e. Verification of records (using statistical quality control) prior to transfer to storage.

5. The problems of creating a complete national bibliographic data store should be studied. This would involve determining the best means of obtaining standardized records for bibliographic items not represented in the Library of Congress record set. The study should also investigate the feasibility of establishing a true national union catalog by recording holdings of American libraries in the machine-readable data store.

6. If the foregoing conclusions and recommendations are accepted:

   a. An implementation committee should be formed to investigate the sources of funds for the following tasks:

      (1) Development of a detailed design of a system in terms of hardware, software, procedures, and administrative organization. This should include consideration of the adaptability of programs of the MARC system and the proposed hardware/software configuration for the LC Card Division mechanization project.

      (2) A pilot project to test the proposed conversion system. Ideally, it would cover the highest priority material (English language records, 1960-1968).

      (3) Long-term operation of the conversion effort.

   b. If funds are procured, a project should be established to carry out the developmental work as quickly as resources permit.

Chapter 3

USES OF CONVERTED BIBLIOGRAPHIC DATA

A prime reason for converting catalog records to machine-readable form is to achieve greater flexibility in manipulating the data. This flexibility will facilitate searching and retrieval; it will lessen the effort of updating the records; and it will contribute to production of a wide variety of cataloging products (cards, book catalogs, special lists, book labels, etc.) Although initially most of the applications will be along traditional lines, computerization of cataloging data should give an added dimension to bibliographic control that may materially alter familiar patterns of use. Since it is beyond the scope of the RECON study to make a detailed exploration of the potential of machine-readable cataloging data, however, this chapter is limited to a general discussion of some of the possibilities.

The conversion of current cataloging records to machine-readable form satisfies needs related to the processing of current acquisitions but, by themselves, current records would not fill the needs of full-scale searching and retrieval. If a data base of machine-readable catalog records is built solely in terms of current and future cataloging output, libraries will have to face the consequences of having a dual system:

part machine, part manual. In practice this means that searches for known items and retrieval of records by subject would often be handicapped by uncertainty as to the proper file to approach and the necessity of using both files.

Library acquisitions do not follow a straightforward pattern that insures obtaining imprints only in the year of their publication. Therefore, in considering potential uses of retrospective bibliographic data in machine-readable form, it should be emphasized that the term "retrospective" has two quite different connotations when applied to catalog records. In the most obvious sense, the term applies to the records for materials already acquired and cataloged for the Library's collection. When this is true, the records can be termed "true retrospective" records. In another sense, however, it applies to catalog records needed for materials published in previous years but currently being acquired and cataloged by a library. Such records may fill a "current retrospective" need. It follows then that the type of application (acquisitions, union catalog, etc.) and the characteristics of both the existing collection and current acquisitions will determine the most useful data base for a given library or library system.

The ability to search existing holdings by machine to avoid ordering unwanted duplicates and to verify a requested item against a reliable data base would be an obvious boon to the acquisition process of any library. To obtain the maximum benefits, a library should have its entire file in machine-readable form. Otherwise, some proportion of the searches will have to be made in both the manual and machine files. This might not be troublesome in a scientific library that acquires virtually no retrospective items because of the high rate of obsolescence of published material in its field. For such a library, time would take care of the problem of dual files. A general library could not anticipate such a simple solution to the problem. Unless its retrospective records were converted it would have to maintain a manual file indefinitely.

Availability of an extensive body of machine-readable bibliographic records would facilitate catalog production and maintenance in all kinds of individual libraries and library systems. Catalogs of an entire library system could be duplicated for branches or departmental libraries. Catalogs that are deteriorating or damaged could be rehabilitated as required and the integrity and security of this major bibliographic tool could more readily be preserved. Catalogs could be updated from changes made to the central record, so that, for the first time, it would be possible for many libraries to keep abreast of changes in descriptive cataloging, subject analysis, and classification.

The availability of converted retrospective bibliographic data would promote uniform standards of classification, descriptive and subject cataloging. Individual library catalogs could be matched to the standard data base to provide union catalogs or system-wide catalogs. Centralized services--acquisitions, production of ready-to-file catalog cards or book catalogs and of book preparation materials (bookcards, pockets, spine labels)--would be more acceptable and generally of better quality if based on retrospective as well as current LC records in MARC format. Commercial services would likewise benefit from such a data base, and could provide complete bibliographic "packages" for libraries which prefer purchasing such services in contrast to entering into cooperative systems or performing the work in-house.

The retrospective data base would also be a source of records for the control of circulation, interlibrary loans, and the rotation of materials among branches of a system. Usually these purposes could be served by a briefer record than would be needed in other applications but there would be no difficulty in abbreviating a standard record if that were desirable. The MARC II format offers great flexibility in selecting data for specially tailored needs.

Automated circulation, acquisitions, cataloging, and interlibrary records could also be analyzed by type of material, subject, language, date, and other characteristics to provide the kind of management data that is so conspicuously lacking in libraries. Such information is needed for planning acquisitions, new buildings, departmental or branch collections, storage space, stack space, work load and staff projections, networks, and many other facilities and services.

The potential applications of converted bibliographic data extend far beyond assistance and cooperation in technical services, data-processing operations, and provision of management information. Substantial benefits could be derived from improved access to bibliographic information. For example, in retrieving catalog records, it should be possible to use subject headings and descriptive information (e.g., language of the text, imprint date) together to reduce the user's effort in a way that it is impossible in present-day catalogs. Using these and other techniques, the machine-readable data base should provide the means of producing special bibliographies that would be far too costly and time-consuming to prepare manually. It is also likely that such bibliographies would be more accurate and exhaustive than those obtainable by human effort. Given the proper hardware and software, the variety of uses of machine-readable cataloging data would be limited only by the imagination of the user.

The provision of new and highly flexible records coupled with greatly expanded file access is likely to stimulate a variety of applications as yet unforeseen. The research questions and/or programs that follow from the existence of a national bibliographic store may be:

1. Consideration of the long range future of the local library catalog.
2. Replacement of the present "all or nothing" approach to bibliography by a graded series of bibliographic records with access time and completeness varying inversely with cost.
3. Rapid dissemination of preliminary records to be replaced later by more complete records.
4. Investigations of users' interactions with large data bases in a variety of environments and styles of presentation; i.e., new and different card files, possibly with different card designs and different file organization; book catalogs;

on-line, interactive searching with and without libraries as "negotiators."

5. Construction and testing of file organization models in a real world environment. It is conceivable that more than one mode of file organization might be developed as a function of differences in file activity, the nature of various entries, or the characteristics of different inquirers.
6. Evaluation of the role of diacritical marks, graphical representations of nonroman alphabets, and vernacular search terms from the viewpoint of international application of machine-readable bibliographic data.

The standardization of the bibliographic record and of its machine format would make possible the transmission and sharing of information among libraries to an extent never before possible. If large files of retrospective records existed, union catalogs either in book form or accessible by terminal could be used to locate materials in a region. On-line retrieval from a bibliographic center or region would also be a possibility but many problems must be solved before it can become a practicality.

Interlibrary cooperation could take many forms, from improved interlibrary loan and cooperative acquisitions programs to elaborate networks utilizing the latest computer and communication technology. All of these advances would depend on access to information beyond the individual library. A program for the conversion of Library of Congress retrospective records to machine-readable form could extend logically to development of a true national union catalog, listing locations of all titles held by American libraries. This possibility is explored in the next chapter. If feasible, it might provide effective national bibliographic control for a true national library network and pave the way for international bibliographic control in combination with the National Program for Acquisitions and Cataloging.

Chapter 4

MASTER DATA BASE

A. Factors Affecting Choice of a Data Base

The selection of a master data base of retrospective catalog records for conversion must take into account the factors of (1) duplication in whatever data base is chosen and the collections of prospective users, (2) acceptability of the data base with respect to bibliographic accuracy and completeness, and (3) forms of material to be excluded.

1. Duplication

Studies of U. S. library collections show that there is considerable duplication (see appendix A) and a recent study indicates that the extent of duplication is increasing. In general, the larger the library, the more likely it is to include the holdings of other libraries, and the more likely it is to own works that other libraries have not acquired. Specifically, the Library of Congress was shown to hold 80.3 percent of the titles held by 11 regional catalogs in 1942. More recent data show that over 50 percent of reports to the post-1956 National Union Catalog are on LC cards, notwithstanding the fact that the criteria for contributing to NUC reduce reporting in categories of material in which extensive duplication is known to occur (e.g., standard U. S. imprints).

These studies constitute a strong argument for focusing the conversion effort on the largest available catalog. It would provide the greatest coverage of titles held by other libraries and at the same time would include many titles not held by any other library. The largest catalog in North America is the National Union Catalog. The LC Official Catalog ranks next, although it is possible that two other research library catalogs may be of comparable size. As will be shown, however, size is not the sole criterion for selection of a master data base.

2. Bibliographic Accuracy and Completeness

To be of maximum usefulness, a national data base should meet an acceptable standard of bibliographic accuracy and completeness. Even allowing for the fact that older LC catalog records have not always been changed as new policies and new cataloging rules have been adopted, few libraries have adopted standard cataloging rules as completely or applied them as consistently as the Library of Congress. The lack of uniformity in the cataloging practices of other U. S. libraries is revealed by striking variations in entry among reports to the National Union Catalog (see appendix A). Thus, the wide dissemination and acceptance of LC cataloging in the form of cards and book catalogs gives it the status of a national standard.

3. Exclusion of Certain Forms of Data

Serials have been excluded from the present study because they are to be converted by the National Serials Data Program of the three national libraries and thus consideration of them in the present study would be redundant. Moreover, the survey of libraries engaged in or contemplating conversion (see appendix B) revealed that many of them were concentrating on monographs. The consultants interviewed (appendix C) were also in favor of focusing on monographs.

Nonbook materials have been excluded from the study for much the same reason. It is the opinion of the working task force, corroborated by the consultants interviewed, that despite the importance of nonbook materials, monographs should have priority for a national data store. In addition, formats for machine-readable records for these materials have yet to be developed. Only after a list of data elements has been agreed on and content designators developed can a standard data base be created.

B. Consideration of Existing Files

1. Library of Congress Official Catalog

The LC Official Catalog is the most suitable choice of the master data base with respect to the completeness, accuracy, and quality of the bibliographic information it contains. Although no comparative studies are available, it seems doubtful that any other library can match the LC record for keeping its catalog up to date.

A study of the extent, method, and types of changes in LC catalog cards is reprinted as appendix E to this report. The study shows that in random samples of cards produced over the last 30 years the average percent of records changed varies from 4.5 percent after one year to 41.9 percent after 30 years. The data elements most frequently changed are subject headings, with added entries and main entries also ranking high.

There are obstacles to using the Official Catalog as the file to be converted. First of all, the name portion of the catalog contains about 12 million cards, including main, added, and subject entries; name authority cards; series treatment cards; and other types of control records. Thus, it would be time-consuming and costly to search this file for all or part of the four million discrete catalog records produced by the Library of Congress since 1898. Second, the master records themselves frequently contain so many additions and changes that they would be difficult or impossible to use in almost any conversion process. The best way to overcome these obstacles would be to first convert the LC Card Division record set (see next section) and then to update the resulting machine-readable records by comparing them with the master records in the Official Catalog. The proposed procedure is described in detail in the following chapters.

2. LC Card Division Record Set

The record set of the Library of Congress Card Division consists of a master copy of the latest revised reprint of every LC printed card, arranged by card series and, within each series, by card number. The fact that the record set is subdivided by card series and can be segregated into specific time periods makes it a tempting candidate for conversion. Not only can a specific time period be selected for conversion (e.g., the last 10 years) but also periods when different cataloging rules and practices were in effect can be readily segregated for special treatment as necessary. Finally, the records, which are clean and legible, appear only once in the file for each bibliographic item.

The primary disadvantage of the record set from the standpoint of conversion stems from the fact that only certain types of changes in cataloging cause the record to be reprinted. Revised reprints result primarily from changes in main entry, title, or other elements necessary for correct identification of the book. Changes in added and subject entries, contents notes, and classification numbers are typed or handwritten in the Library's own catalogs and remain in this form unless the card is reprinted for another reason. Since these changes do not appear in the record set, a data base produced from this source alone would be seriously out of date, and the burden of updating added author and subject entries would be placed on the user libraries. This would mean changing the same record many times in many places. Apart from the repetitious labor involved, this approach would be unsatisfactory because local updating would not always be done in a standard way.

3. National Union Catalog

The National Union Catalog contains an estimated seven million titles in addition to the approximately four million LC records, and thus consitutes a more complete data base than the LC catalog. The types of publications that figure most prominently among titles not covered by LC cards include: dissertations; state and local publications; analytics;

foreign language titles; and editions that LC catalogs as copies. These categories do not reflect the titles most duplicated among various library collections and therefore most in demand from a national data base.

The variation in entry reported to NUC for the same title has already been mentioned. There is no effective standardization in the reports as received by NUC, other than those reported on LC cards, and the NUC editing operation attempts only to check the main and added entries for conformity to established LC form. The body of the card and the subject headings (if present) are not edited in any way. It would, therefore, be impossible to create a data base conforming to any acceptable standard of accuracy and uniformity from the National Union Catalog. The desirability of including non-LC cataloged items for an eventual true national data store is discussed in section F of this chapter.

4. Library of Congress Shelflist

The LC shelflist has been suggested as a desirable source file for conversion. This approach, based on experience with the Harvard shelflist conversion project, favors conversion by subject groups. The pros and cons of a subject approach are discussed in section C and appendix C.

The overwhelming disadvantage of this method as far as the LC shelflist is concerned stems from the composition of that file. It contains a mixture of temporary, incomplete, and printed records with essentially no corrective changes beyond revision or updating LC class and book number. Nor are the cards legible enough to be microfilmed to provide a readable guide to locating the master records in the Official Catalog.

Various languages, alphabets, and different eras of cataloging rules are not easily separated in the shelflist.

C. Approaches to Conversion of the Master Data Base

The choice of the monographic records in LC Official Catalog as the master data base for conversion leaves unresolved the problem of how such an immense a task could be undertaken. Even if the goal is total conversion, priorities must be established because, under the best circumstances, the time required for the job must be reckoned in years. As a practical matter, therefore, it is essential to define subsets of the file to insure that maximum benefits can be obtained for the effort expended.

Portions of the master data base can be selected for conversion on the basis of (1) subject, (2) special bibliographies, (3) date, (4) language, and (5) on-demand requests.

1. Subject

A subject approach to conversion has the appeal of providing packages that, superficially, can be defined with a certain amount of precision. In practical terms, however, a priority scheme based on subjects is highly impractical because the LC shelflist is not usable either as the master data base or as a record for initial conversion with subsequent update from the Official Catalog. Furthermore, the appeal of the subject approach appears to be limited since the library survey (appendix B) showed that only a small number of libraries actually involved in conversion were concentrating on specific classes, and that these conversion efforts ranged over many subject areas with little duplication.

2. Special Bibliographies

In the opinion of several of the consultants interviewed, the best return on funds expended for conversion and the greatest utility would be attained by converting such published lists as Book for College Libraries (BCL), Books in Print, etc. Procedural problems of getting from the lists to the up-to-date LC record are a major deterrent to this approach. An effort now in progress to convert BCL for "current retrospective" use is reported in appendix C. Putting the catalog records for a specific list in machine-readable form is primarily beneficial to users who base their acquisitions on the list. Other users seeking machine-readable records for specific titles would have to determine whether the title appeared in the printed booklist before requesting the record or face the likelihood of a large number of unsuccessful searches.

3. Date

The consultants agreed that conversion of records produced in the last five to ten years should be given first priority. The library survey also reported that a majority of libraries actually involved in conversion were concentrating on specific time spans, mostly within the period of the last ten years. Among libraries contemplating conversion, fewer plan to impose time limitations, but when they do, the period 1960 to date predominates. Reverse chronological conversion of the master data base is easily accomplished because the LC record set is arranged by card-number date and thus falls into manageable groups.

4. Language

An overwhelming majority of the consultants favored conversion of English language records first. Results of the library study showed less than half of the libraries involved in conversion were concentrating on specific languages, but, of these, almost all were concentrating on English language works. The disadvantage of categorization of records in the master data base by language is that it must be done almost entirely manually.

5. Demand

Similarities between a service to distribute machine-readable data for retrospective records and the present Card Division service suggest that it might be reasonable to convert older records on demand. In this method, conversion would be stimulated by actual requests from other libraries. If the evidence of duplication is valid, this method would gradually produce a data base capable of serving a large proportion of user needs. It would seem to have the advantages of eliminating unused records from the conversion effort and accommodating a range of languages.

On the other hand, conversion on demand has many disadvantages. First, it would sacrifice many of the efficiencies of systematic conversion which allow orderly organization of the work flow. In practice, it would lead to the establishment of interior priorities as to which requests should be given preference. Otherwise, a strict "first-in-first-out" flow could result in the conversion of records in minor foreign languages causing serious delay to the conversion in titles in English, Second, and most serious, is the fact that the heterogeneous character of the resulting data base would make it very difficult to predict whether a given title had been converted. Thus, many searches against the machine data base would be fruitless. Since the analysis of a hypothetical on-demand service (see appendix H) indicates that demand searches would consume costly processing time, it would be highly doubtful whether the system could afford a high proportion of unsuccessful searches. Systematic conversion by language and data overcomes these difficulties to a large extent. In view of the coverage of Library of Congress cataloging, there would be a high probability of satisfying a request that fell within the scope of a data base of, say, English language records since 1960.

Finally, as far as the Library of Congress is concerned, the on-demand strategy would be of doubtful value in building a data base for retrieval since it would have the effect of limiting the coverage to that part of the LC collections held by other libraries.

Despite the disadvantages of on-demand conversion, it might be possible to combine this strategy with systematic conversion by language and date, if this could be done without too great a reduction in efficient processing. This possibility should be explored to meet the anticipated needs of the LC Card Division.

D. Priorities

The conversion of currently produced catalog records did not seem originally to be a concern of the present study. It became apparent however, that the disadvantages of adding to the already heavy load of retrospective records made it urgent to move as quickly as time, staff, and the state of the art allow toward the goal of conversion of all current cataloging to machine-readable form. It was logical also to conclude that no effort should be expended on retrospective conversion of any subset of the total body of catalog records unless the MARC Distribution Service was converting current records in that category.

For various reasons it is not possible to predict when the Library of Congress will be able to convert all of its cataloging output on a current basis. To provide benchmarks for estimating the workload of conversion of retrospective records, however, the following starting times were used for each major category:

Category	Beginning date
Romance and German languages	July 1970
Other roman alphabet languages	July 1971
Nonbook materials	July 1971
Slavic languages	July 1972
Other nonroman alphabet languages	July 1973

The beginning dates were staggered in the expectation that the expansion of the MARC Distribution Service would be phased to allow an orderly buildup of staff. The schedule also allows time for the resolution of conversion problems such as processing nonroman languages. It should be kept in mind that these dates were established for purposes of calculation in the RECON study. They do not represent operational decisions by the Library of Congress. Appendix D gives detailed tables of the workloads for retrospective records and anticipated cataloging production through June 1976.

On the assumption that the Library of Congress may be able to initiate conversion of current cataloging for the various categories according to this schedule, the following groups of retrospective records might be considered for conversion.

	Category	Time span	Number of records
1.	English language	1960-March 1969	386,000
2.	Romance and German languages	1960-June 1970	381,000
3.	English language	1898-1959	1,728,000
4.	Other roman alphabet languages	1960-June 1971	137,000
	Nonbook materials	1960-June 1971	157,000
5.	Slavic languages	1960-June 1972	225,000
6.	Other nonroman alphabet languages	1960-June 1973	256,000
7.	Romance and German languages	1898-1959	698,000
8.	All remaining catalog records	1898-1959	682,000

It is recommended that first priority be given to conversion of English language monographic records back to 1960. The evidence of this report shows overwhelmingly that these records will satisfy the largest proportion of the needs of prospective users. Second priority should be given to conversion of Romance and German language records back to 1960 because they serve an identifiable need in academic and research libraries. The third priority should be accorded to English language records back to 1898 (the earliest LC printed cards). Completion of this phase of the conversion effort would provide a complete span of readily definable catalog records from which all types of libraries could build data bases that should satisfy the vast preponderance of requests for information retrieval. While it must be acknowledged that all of these categories include records of questionable interest and utility, it was felt that the high cost of identifying these marginal records would largely offset any savings to be realized by eliminating them from the conversion effort.

When records in the first three priorities have been converted, further steps should be considered in the light of user needs and technological capabilities at the time. It did not seem realistic within the constraints of the present study to assign absolute priorities to the remaining categories. Defining and quantifying them, however, provides a foundation for further study and consideration.

E. Strategy for Conversion

In summary, monograph records from the Official Catalog are recommended as the master data base. This data base would best be created in a two-step process by converting the LC record set and subsequently updating the record from the Official Catalog.

Since the record set is an active working file for the Card Division, it cannot be used directly as input for conversion. The essential features of the proposed approach would involve sorting the record set into categories of conversion priority. The groups of records, once microfilmed, would be reconstituted into the original record set. The microfilm data would be converted according to priority, and the results of the conversion would be matched against the corresponding records in the Official Catalog. When appropriate, the converted records would be revised to correspond to additions or changes found in the Official Catalog.

Since libraries now accept records that are not entirely up to date when they obtain cards from the Card Division, the question naturally arises, "Why can't the machine-readable records be of the same quality?" Several answers may be made to this question:

1. It can hardly be argued that the present limitations on the currency of the catalog cards are desirable.
2. In the present situation a library generally obtains only a few older records at a time. When they are merged in its catalog, their headings must be reconciled with those already present. In the future, if a library is engaged in wholesale conversion to machine-readable catalog records, it may be able to accept Library of Congress headings without change provided they are consistent and up to date.
3. Even if other libraries were willing to accept uncorrected records, it would be inconceivable that the Library of Congress would accept a machine-readable data base of lower

quality than the Official Catalog. Since the records would have to be updated for that purpose, it seems reasonable to allow all potential users to share the benefits.

4. Any consideration of using the products of a retrospective conversion project as a basis for a national bibliographic store necessarily depends on the records being of the highest quality obtainable.

In this connection, it should be noted that the conversion project would not result in static records. These records would be subject to change at the same rate (approximately one percent of the total data base each year) as now occurs in the LC Official Catalog. Therefore, the value and integrity of the data base could be preserved only by making these changes when they became known. Any other course would lead to the gradual obsolescence of the file.

F. Considerations Regarding a National Data Store

One further possibility, which must be outlined even though it remains imprecise and hypothetical at present, is the accumulation and use of machine-readable bibliographic records in a national data store analogous to the National Union Catalog. It would provide a repository for converted titles from all libraries as well as a record of their holdings.

The arguments for implementation of this concept are basically the same as those which led to the creation of the National Union Catalog and its eventual publication in book form. The arguments are enhanced for a machine-readable data base by the increased ease in manipulation and speed in transmission of information which such a system may offer.

The proposal of such a scheme adds a whole new set of problems to those already present in plans for retrospective conversion of the bibliographic records of the Library of Congress. To begin with, the question of centralized versus decentralized conversion and reporting reappears in a new guise. The National Union Catalog, with records of holdings, already exists. If the records in this catalog were to be converted along with information about titles held by specific libraries, a number of decisions would have to be made on accepting or revising known types of inaccurate, incomplete, and obsolete data, as well as data known to be missing. These include (1) withdrawn items for which no notification has been given NUC, (2) holdings unreported because they belonged to a library not participating at a given period of time, or because of the limited number of holdings accepted from a given geographic region by NUC, and (3) incomplete or inconsistent bibliographical data reported for the same item by different libraries with regard to choice of entry, form of heading, use and application of subject headings, and the like. The resolution of these problems would constitute an enormous task. The alternative of converting the data in the National Union Catalog exactly as it stands, although perhaps easier to execute, would lead to a product of considerably lesser utility.

Another alternative is based on the conversion of one of the data bases described in this report. Once the basic store of data was converted, whether it be the Official Catalog of the Library of Congress, the full record set in the LC Card Division, or a block of records such as all English language titles cataloged during the past ten years, other libraries engaged in the process of converting local holdings might participate in a national plan for reporting converted records to incorporate them with the basic store. A fundamental requirement of any such plan would be adherence by reporting libraries to at least minimal standards prescribing (1) content of any particular bibliographic record reported, and (2) content designators for those data elements reported. A prototype of a minimal record appears in appendix F as level 3.

A number of implications follow from dependence on a reporting plan alone in contrast to conversion of the National Union Catalog. By definition, the plan would be limited to only those libraries with the capability of converting bibliographic data. During the next few years, it is unlikely that a large number of libraries will have this capability. Indeed, it is to be expected that a number of smaller libraries with significant research collections in special fields will not be in a position to convert bibliographic data for many years. This procedure for adding local holdings to the national data store would thus depend on a factor almost completely unrelated to potential utility; that is, the development and implementation of automated bibliographic systems and the adoption of reporting procedures at particular libraries throughout the country. Reliance upon local reporting would not guarantee that the national data store truly or even significantly reflected the bibliographic holdings of the library community. It will be necessary, therefore, to find other means for obtaining up-to-date information about the holdings of libraries to be represented in the national data base.

A national bibliographic data store should naturally incorporate currently cataloged titles. This is a corollary to the extension of the MARC Distribution Service as soon as possible to all current Library of Congress cataloging data to prevent the further accretion of "retrospective" bibliographic information not in machine-readable form. Any plan to create a computerized NUC should include procedures for adding to currently produced MARC records the locations reported to the National Union Catalog by libraries throughout the country. While the actual addition of locations in machine-readable form involves few theoretical problems, the timing raises considerations that will need careful attention if this information is to be distributed to regional centers.

The MARC Distribution Service exists to distribute cataloging information that will facilitate the organization of current library acquisitions. Speedy execution of this task is essential to its success. It follows then that the service cannot also be the vehicle for distributing information about libraries that hold titles included on the current tapes because this information is usually not available for many months after the item has been cataloged by the Library of Congress. Some other means must be found to distribute holdings information to those regional centers that may be involved in the creation of union catalogs of machine-readable data.

Even more difficult will be the establishment of ground rules for the reporting by local libraries of titles which, when cataloged locally, have not been acquired by LC and/or included in MARC. Will the title be acquired by LC? Will it be included in MARC? How will matching of machine-readable reports be done in the central file if conflicting data are reported by two or more libraries?

The present report will not present either a detailed scheme for creation of a national bibliographic data store with holdings, nor a cost estimate for the accomplishment of this task. To provide this information would require a study complementary to the present one and of the same or greater magnitude. Such a study would be premature before some of the proposals and recommendations outlined in the present report have been acted upon. If conversion of retrospective bibliographic data becomes a reality, however, its fullest benefits will be realized only if information giving nationwide holdings is made available through conveniently accessed means.

Chapter 5

TECHNICAL ALTERNATIVES: MACHINE CONSIDERATIONS

A. Basic Assumptions

1. Introduction

Chapters 5 and 6 deal with the machine and manpower considerations of the technical alternatives[1] that were analyzed during the course of the study. Many of the concepts and assumptions discussed in both chapters are described in detail in the appendixes. Chapters 5 and 6 are highly interdependent and, in addition, they assume that the reader is familiar with certain concepts, terminology, and basic assumptions.

It is necessary, therefore, to define these terms and assumptions (1) to avoid duplication of definitions in chapters 5 and 6, (2) to clarify the contents of the two chapters for the reader without requiring him to refer to an appendix to understand terminology, and (3) to explain some of the basic assumptions underlying the technical alternatives.

---

[1] In this study the term "technical alternative" embraces all facets of the conversion process after the selection of the data base through the quality-control step prior to storage. Initially, the data would be recorded on magnetic tape. In the operating system the data would be in a random-access mass storage device. The term "conversion method" is also used to refer to the same process.

2. Staff Complement

To determine staffing requirements and elapsed time for each alternative data base, some hypothetical conversion rates had to be assumed. A basic premise of the study was that all aspects of conversion should be assessed realistically. Therefore, the size of the staff for a given conversion method could not be so large as to make staffing impractical. The RECON Working Task Force felt that a staff complement of about 100 people was realistic and it was calculated that approximately this many people could implement a conversion effort of 10,000 titles per week regardless of the technical alternative chosen. The tables included in chapter 6 were constructed on this basis. It should be noted that the staff complements in the tables can be used as a base for calculating the production of different numbers of people or different rates of production. This could be done to increase or decrease the time required to convert any particular data base.

3. Editing

The term "editing" has been used rather broadly in this report and sometimes encompasses several distinct processes. Actually editing can be defined as the process of applying tags, delimiters, and subfield codes (content designators) and adding certain fixed field information (language code, main entry in body of the entry, imprint date, etc.) to the record. In the analysis of staff requirements for the various technical alternatives, the process of editing includes the original editing, proofing for completeness and accuracy, and correcting errors. Although the proofing and correction processes occur at a later stage, they are considered part of editing since each technical alternative assumes the same people are performing all functions (based on MARC I and MARC II procedures).

The process of editing the record prior to input is called pre-editing; the process of correcting the record after proofing is called post-editing.

The process of pre-editing can be performed at three different levels of completeness:

Full editing assumes the human editor has assigned all content designators and the machine processing does not include a format recognition program (see below).

Partial editing assumes the human editor has assigned some content designators and the machine processing does include a format recognition program that analyzes the record and assigns the remaining content designators. The content designators assigned by the human editor in partial editing are called cues to the format recognition program. The content designators would aid the machine analysis and increase the accuracy in the format recognition program.

No editing assumes that there is no pre-editing process and the machine processing includes a format recognition program that assigns all content designators by analyzing the character strings.

The terminology "full editing, partial editing, and no editing" is used when describing editing as a process. When the resulting record is being described, the adjectival forms of all three are used: fully edited record, partially edited record, and unedited record.

The post-editing process is always the same. Regardless of the types of pre-editing the record has received, post-editing will add or change content designators or characters in the bibliographic description itself (i.e., misspelling, keying errors, etc.).

4. Format Recognition

Format recognition is a function performed by a computer program. The function may be defined as the analysis of the data in a machine-readable record and the automatic assignment of content designators (tags, delimiters, and subfield codes) and coded information (fixed fields) making explicit what is implicit in the textual information (language codes, form of content, etc.).

Format recognition is not applicable to fully edited records. The term "fully edited" implies that a human editor has already performed the function. Partially edited records have received some treatment by a human editor. The machine uses the information provided by the editor as cues to complete the assignment of content designators and fixed fields. Unedited records are input directly into machine-readable form without any manual editing and the format recognition program attempts to assign all the content designators and fixed fields required. An extended treatment of the concept of format recognition appears in appendix G.

5. Levels

The concept of levels of records and its development for this study is explained in appendix F. A machine format for recording of bibliographic data and the identification of these data for machine manipulation is composed of a basic structure (physical representation), content designators (tags, delimiters, subfield codes), and contents (data elements in fixed and variable fields). Although the basic structure should remain constant, the contents and their designation is subject to variation. For example, a name entry could be designated merely as a name instead of being distinguished as a personal name or corporate name. When a distinction is made, a personal name entry can be further refined as a single surname, multiple surname, or forename. Likewise, if a personal name entry contains date of birth and/or death, relationship to the work (editor, compiler, etc.), or title, these data elements can be identified or can be treated as part of the name entry without any unique identification. Thus individual data elements can be identified at various levels of completeness.

The MARC II format[2] for current cataloging data has been defined as level 1. This constitutes the most complete record and assumes that the physical book was inspected during conversion. Level 2 has been defined as a MARC II record prepared without consulting the original book. This may mean that some data elements may not be supplied. As can readily be seen, the gradual elimination of data elements and content designators would produce formats at different levels of completeness. Thus, level may be defined as the completeness of the record in terms of content and/or content designators.

---

[2] U. S. Library of Congress. The MARC II format, a communications format for bibliographic data. Washington, D. C., 1968.

B. Description of Possibilities

After analyzing the alternative data bases (see chapter 4) it was necessary to develop several conversion methods so that unit costs/record could be calculated for input equipment and staff requirements for each conversion method. The cost of the computer system (including both hardware and software) needs to be calculated only once because it is unaffected by the choice of the data base or the conversion method. The design and implementation costs for the necessary software would remain constant regardless of the number of records to be converted. For design purposes it was assumed that the computer configuration should process a data base of from one to five million records, enough for any of the data base alternatives under consideration. The costs of the mass storage devices would depend on the number of records converted. Therefore, these costs would be affected both by the total number of records converted and by time. It would not be necessary to procure the total number of storage devices at the onset of the conversion effort. They could be added as the data base grows. In the design of the machine configuration (see appendix H), both selection of hardware and software specifications were based on the time frame 1970-1976 and thus they reflect what is available today.

Since it was decided that the LC record set (updated from the Official Catalog) was the best file for conversion regardless of the data base or the conversion method finally recommended, the selection process would remain constant for each technical alternative. Therefore, it was necessary to compute the costs of selection and the hardware/software configuration only once. Attention was then focused on evaluating the advantages and disadvantages of various conversion methods and determining their costs.

Figure 5.1 illustrates the alternative conversion methods. Each lettered alternative (A-J) represents a form of editing (no editing, partial editing, full editing) using a different keying device (for this discussion, direct-read OCR is classified as a keying device).

Alternative	Form of pre-editing	Input device
A	None[3]	Direct-read OCR
B	None	Magnetic tape inscriber
E	Partial	
H	Fully	
C	None	OCR font typewriter plus OCR
F	Partial	
I	Fully	
D	None	On-line typewriter
G	Partial	
J	Fully	

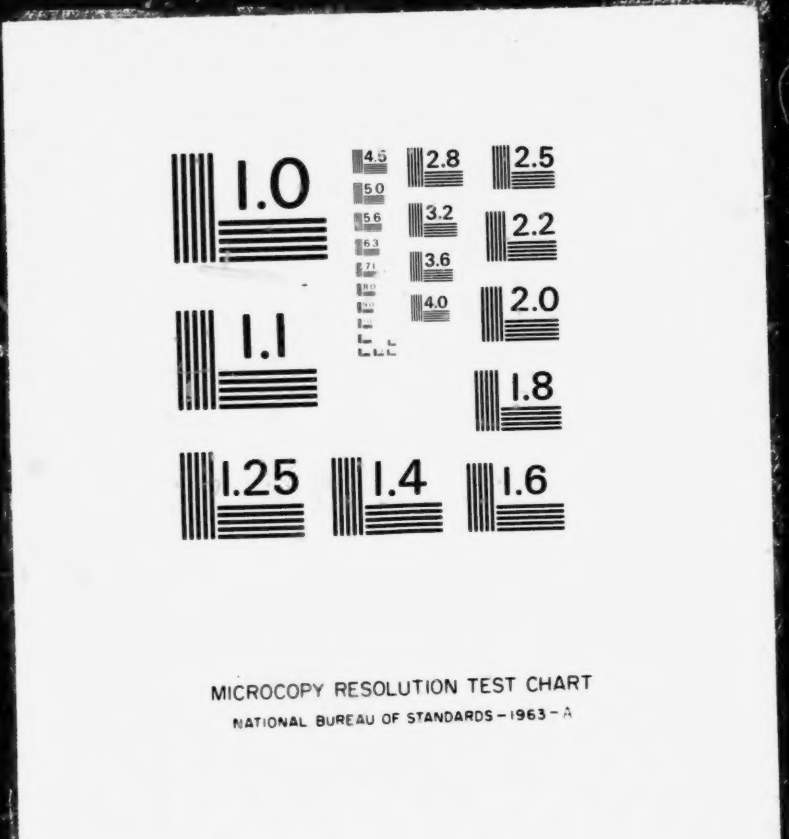
---

[3] The alternative using direct-read OCR does not lend itself to pre-editing.

2 OF 4

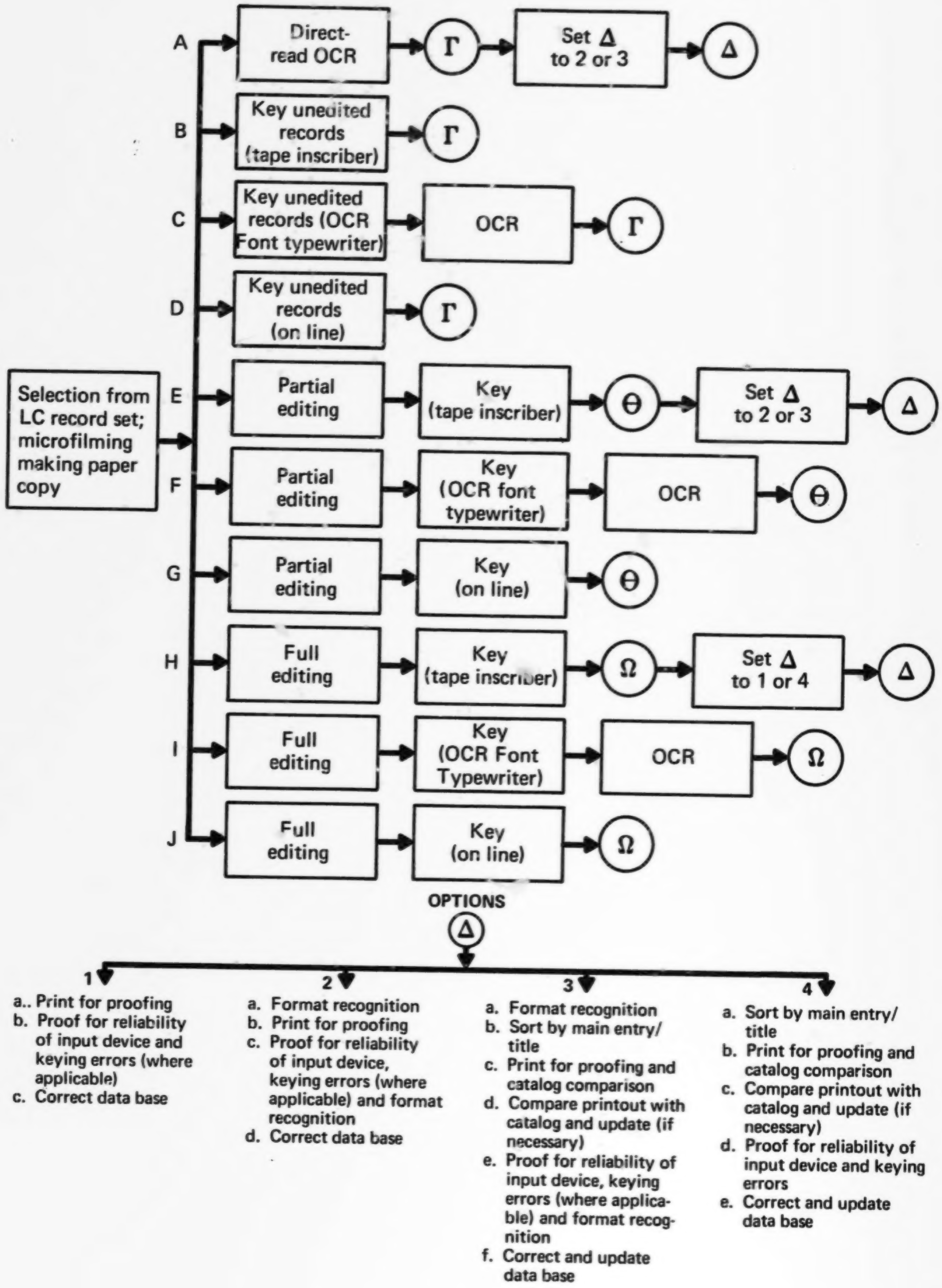
ED

032895


1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
4.5
5.0
5.6
6.3
7.1
8.0
9.0

MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

Figure 5.1--Technical alternatives for conversion of LC catalog records to machine-readable form


Selection from LC record set; microfilming making paper copy
A
Direct-read OCR
Γ
Set Δ to 2 or 3
Δ
B
Key unedited records (tape inscriber)
Γ
C
Key unedited records (OCR Font typewriter)
OCR
Γ
D
Key unedited records (on line)
Γ
E
Partial editing
Key (tape inscriber)
Θ
Set Δ to 2 or 3
Δ
F
Partial editing
Key (OCR font typewriter)
OCR
Θ
G
Partial editing
Key (on line)
Θ
H
Full editing
Key (tape inscriber)
Ω
Set Δ to 1 or 4
Δ
I
Full editing
Key (OCR Font Typewriter)
OCR
Ω
J
Full editing
Key (on line)
Ω
OPTIONS
Δ
1
a.. Print for proofing
b. Proof for reliability of input device and keying errors (where applicable)
c. Correct data base
2
a. Format recognition
b. Print for proofing
c. Proof for reliability of input device, keying errors (where applicable) and format recognition
d. Correct data base
3
a. Format recognition
b. Sort by main entry/title
c. Print for proofing and catalog comparison
d. Compare printout with catalog and update (if necessary)
e. Proof for reliability of input device, keying errors (where applicable) and format recognition
f. Correct and update data base
4
a. Sort by main entry/title
b. Print for proofing and catalog comparison
c. Compare printout with catalog and update (if necessary)
d. Proof for reliability of input device and keying errors
e. Correct and update data base

In addition to the major breakdown of A-J, there is a secondary division categorized as 1-4.

The secondary divisions 1-4 cover procedures that might be followed after the data were on magnetic tape as a result of the A-J conversion methods. The following section elaborates the details of the secondary divisions 1-4:

1: The magnetic tape records are printed for proofing against a source document for the reliability of the input device (machine errors) and keying errors (when a keying device is used), and the records then are corrected, keyed, and input to correct the machine-readable data base. Since the records are not processed by a format recognition program, the quality of the resulting record would depend on the type of pre-editing the record had received. For example, if partial editing was performed, the resulting record would be in a format somewhat less complete than level 2 (see appendix F). If full editing was performed, the resulting record would be in a level 2 format. If the record was unedited, the resulting magnetic tape record would be a character string without any explicit identification.

2: The magnetic tape record is processed by a format recognition program and the record is printed for proofing for reliability of the input devices, keying errors where a keying device was used, and reliability of the format recognition program. Corrections are made, keyed and input to correct the machine-readable data base. The records are not compared against the Official Catalog. Again, it must be borne in mind that the success of the format recognition program depends on the amount of editing performed. The performance of the format recognition program directly affects the number of corrections that will have to be made and consequently the number of records that will be recycled during the conversion process.

3: The magnetic tape record is processed by a format recognition program. The file is sorted by 10 characters of the main entry.[4] The records are printed, compared against the entry in the Official Catalog, and updated, if necessary. The records are proofed for reliability of the input device, keying where a keying device was used, and for reliability of the format recognition program. Corrections are made to the record and both the corrections and changes from the comparison with the Official Catalog are keyed and the machine-readable data base corrected and updated.

4: The file is sorted by 10 characters of the main entry. The records are printed, compared against the entry in the Official Catalog, and updated, if necessary. The records are proofed for reliability of the input device and keying errors. Corrections are made to the record and both the corrections and the changes from the comparison with the

---

[4] The source data were originally taken from the LC Card Division record set. This file is in chronological order by year and within year by sequential number (LC card number). It is necessary, therefore, to sort the file by main entry to facilitate comparison with the Official Catalog.

Official Catalog are keyed and the machine-readable base corrected and updated.

Although there are 40 combinations of the 10 major conversion methods (A-J) and the four secondary options (1-4), figure 5.1 presents only the 20 possibilities that seemed realistic. In the group A-D, options 1 and 4 were excluded because a record without any editing or format recognition would be an undifferentiated character string of bibliographic information. Partially edited records require format recognition to bring them up to level 2. Therefore, options 1 and 4 were excluded from E-G. Because it would be redundant to apply format recognition to records that were fully edited prior to input, options 2 and 3 were excluded from H-J. Figure 5.1 lists the remaining possibilities: A2, A3, B2, B3, C2, C3, D2, D3, E2, E3, F2. F3, G2, G3, H1, H4, I1, I4, J1, and J4. In the subsequent analysis of these 20 conversion methods they are referred to by this terminology.

C. Input Devices

During the initial phases of the study, several input devices were considered and, for a variety of reasons discussed below, several devices were excluded from the technical alternatives. The decisions made in this phase were made on technical grounds only, not on a comparison of cost.

1. Keyboard to Card (Keypunch)

The lack of hard copy for verification as a result of punching, the limitation of the character set on the keyboard, and the limitation of the 80-column card for punching variable-length bibliographic data were considered to be serious drawbacks and this method was excluded.

2. Keyboard to Paper Tape (Paper Tape Typewriter)

This device does produce hard copy as a byproduct of punching and has a keyboard with a larger character set than a keypunch machine. The mechanical punching mechanism often produces errors, however, and the handling of punched paper tape presents a logistic problem. Since the newer devices (e.g., magnetic tape inscribers) are basically the same type of device without the two limitations (mechanical punching errors and paper tape handling), the paper tape typewriter was excluded.

3. Keyboard to Magnetic Tape (Magnetic Tape Inscriber)

Magnetic tape inscribers are of two types: keypunch to magnetic tape and typewriter to magnetic tape. The resulting magnetic tape is computer-compatible tape in some instances and in others requires a converter to translate from the inscriber output tape to the computer input tape.

Although a keypunch-to-magnetic-tape device affords flexibility in error correction and verification and the output is magnetic tape instead of paper tape, the keypunch has the same limitations described in 1 above and consequently the keypunch-to-magnetic-tape device was excluded.

A typewriter-to-magnetic-tape device has all the advantages of the paper tape typewriter: hard-copy output, a larger character set than the keypunch, free-form input for variable-length data without the associated disadvantages of paper tape output and a mechanical punching unit. Therefore, typewriter-to-magnetic-tape was retained as a possible conversion method for further analysis.

4. On-line Keyboarding (Typewriter)

This device has the same advantages as any other device using a typewriter for input plus the additional feature of not requiring any interim medium such as paper tape, or non-computer-compatible magnetic tape prior to final residence on the system's magnetic tape. Therefore, on-line keyboarding via a typewriter was retained as a possible technical alternative.

5. Typing for an Optical Character Reader (Typing and Scanning)

There are several optical character readers commercially available on the market today that require the use of a typewriter equipped with a special font (shape and form of character produced by the typewriter) and a pin feed for better alignment so the hard copy produced is not skewed causing errors during the OCR read time. The data are typed on a data sheet which is fed through the reader. Each character is interpreted, digitized, and recorded on magnetic tape under program control.

The types of OCR can be characterized as follows:

(1) Devices that can read only a stylized uppercase type font where the typist is required to use special characters to indicate upper- and lowercase, punctuation other than

commas, periods, etc. This limitation causes a decrease in typing speed when the source data are as complex as a bibliographic description.

(2) Devices that are capable of reading upper- and lowercase with extended punctuation.

(3) Devices that are not programmable (minimum program capability wired into the device) and very limited in formatting capability.

(4) Devices that are programmable and are much more flexible in formatting capability.

These characteristics were analyzed and only devices satisfying points 2 and 4 were retained for consideration. Since the typewriter for OCR has the same advantages as the typewriter for all other conversion methods discussed above, typing for an optical character reader was retained as a possible technical alternative.

6. Direct-Read OCR

Direct-read OCR in the context of this study is defined as directly converting the LC Card Division record set into machine-readable form without any intermediate keying devices.

A detailed study performed for the Library of Congress concluded that there is no OCR equipment available today that can directly convert the LC Card Division record set. The principal problems involved with the present equipment are the requirements to be able to read (1) proportional spacing, (2) non-standard fonts, (3) special characters including diacritics, and (4) 3" x 5" cards.

Several manufacturers are developing equipment that looks promising at least for portions of the record set. One manufacturer believes that his equipment would be available by late 1969 or early 1970. The state of the art should be monitored continually to determine where and when breakthroughs are likely to appear.

The lack of a commercially available direct-read OCR capable of handling the retrospective records makes it risky to depend on this method for large-scale conversion. Since, in all probability, developments by manufacturers will be geared to the largest market, it is unwise to anticipate the solution of problems that are beyond present technical capabilities.

Even when an OCR device is available, it will not have the capability to read every character that it may encounter in a record. Two machines that may be available in the foreseeable future require microfilm input. If the quality of the reproduction is poor, the device will be unable to interpret even English letters. In addition, the device will be limited in the number of different characters that it can recognize. It will not be able to read nonroman characters, diacritical marks, mathematical symbols, and other special characters.

To safeguard against digitizing of records with a large number of unread characters, it should be possible to establish a threshhold of the tolerable number of unread characters. If that number were exceeded, the OCR device would reject the record and delete whatever parts it had already read. It has been estimated that, even if the records were preselected so as to maximize the capability of the OCR device, as many as 10 percent would be rejected as unreadable. This figure was taken as the basis for calculating the keyboarding effort required to input these records in the alternative using a direct-read OCR device.

A recent article[5] by a staff member of one of the OCR manufacturers includes the following statements that are directly applicable to the RECON study:

> First, what they record ... By 1975, most OCR applications will involve reading some alphabetic information. There will be a major trend away from the current practice of using retyping and OCR as a conversion method. The move will be to direct reading, which provides the ultimate payoff from OCR. The truly multifont application will be commonplace.
>
> Second, how well they read ... This is now and will continue to be the most important question to be answered in evaluating reading machines. Improvements will be made in readers and in input preparation devices, but many input documents will still be prepared by humans in uncontrolled environments, and the cost of correcting mistakes that get into a computer and of manually handling rejected documents will rise continuously.
>
> Third, how they read ... By 1975, there will be an increased demand for broad flexibility in input formats accepted, and optical readers will have to be capable of performing a substantial amount of on-line computing as a byproduct of the input process.

Chapter 4 discusses the conversion of the LC Card Division record set over a period of years on a priority basis. Because significant advances in the OCR technology can be expected in the 1970's, it is worth considering direct-read OCR as a conversion method for some set of the eight data base alternatives identified. In converting a large data base, many techniques should be considered and no limitation need be placed on the variety employed if a combination of techniques reduces costs. In view of these considerations, direct-read OCR was retained as a possible input device because it might be useful for conversion of some part of the data base that remained to be converted when a practicable OCR capability was developed.

---

[5] Philipson, Herman L., Jr. Optical character readers to play more important role in 1970's. Computerworld, v. 3, February 5, 1969, 4-5.

7. Summary

The input devices considered in the analysis of the unit cost per record for various technical alternatives were (1) direct-read OCR, (2) magnetic tape inscriber (typewriter), (3) OCR font typewriter followed by OCR, and (4) on-line typewriter.

D. Input Costs

1. General Considerations

The unit cost/record figures for the input devices described above were calculated for the transcription of three types of records, each receiving different treatment prior to input; i.e., no editing, partial editing, and full editing. The effect of the three kinds of editing is a difference in (1) total number of characters to be input and (2) complexity of the record to be input. Complexity is measured by the number of content designators (tags, indicators, subfield codes), and the inherent nature of the data itself (for a full discussion of this point, see chapter 6).

Since this study is concerned with records in many languages, allowance was made for a reduced rate of production on input devices using a standard typewriter keyboard because of the complexity of the data. Although there may actually be differences in the keystroke rates for tape inscriber, typewriter with OCR font, and on-line typewriters, they are too slight at this degree of complexity to warrant calculating the separate rates. Therefore, in this study, a uniform rate of 6,600 characters per hour was used for all devices.

Another factor that enters into the calculation of unit cost estimates is the number of characters per record to be input. Based on a statistical study of a random sample of the LC Card Division record set and a count made of the number of characters per record on the MARC II test tape (which includes tags, delimiters, etc.) the following assumptions were made about an average number of characters per record:

Unedited record	325 characters
Partially edited record	412 characters
Fully edited record	500 characters

The character count for a partially edited record was derived by interpolation between the counts for an unedited record and a fully edited record.

The cost of any equipment that an operator uses must take into account the fact that the equipment is not being used for eight working hours a day. Chapter 6 states that all production rates for people were estimated on the basis of an effective working day of six hours.

Therefore, the cost of equipment must be adjusted by an actual utilization factor; in this case 75 percent. All equipment costs were based on a one-shift operation or 176 hours per month.

Some input devices require associated equipment involving a fixed cost that must be prorated over the number of devices actually used. To simplify the calculation of the per-record cost of this associated equipment, whenever an alternative required such a configuration it was assumed that 20 input devices were being used. This assumption was based on an evaluation of the manpower requirements for input discussed in the next chapter. In a few instances, the assumption has the effect of making the per-record cost of the associated equipment different than it would actually be under operating conditions because the technical alternative requires a larger or smaller number of devices.

2. Cost of Equipment

The cost per hour of the equipment for each conversion method is constant regardless of whether the input consists of unedited, partially edited, or fully edited records. In the case of the OCR scanner, however, it was necessary to take these differences into account because of the reading rate of the device.

a. Direct-read OCR

Since there is no commercially available OCR capable of directly reading the Library of Congress printed card, the prices used for cost comparison are based on expected price and rental figures given by the only manufacturer willing to quote a firm price at this time. Read time is also based on projected figures by the developers of the equipment. The quoted rental price was $600 per hour. The projected read time is approximately one card per second or 3,600 records per hour. Therefore, the cost for direct-read OCR is $.167 per record. The cost of the direct-read OCR device on a service bureau basis is assumed to be $600 per hour.

b. Magnetic tape inscriber (typewriter)

Monthly rental	$100.00	
Hourly cost (based on 176 hours per month adjusted for 75 percent utilization factor)		.757/hour
Cost of converter--monthly rental	260.00	
Amortization over 20 tape inscribers	13.00	
Hourly cost (based on 176 hours per month)		.074
Total cost of tape inscriber		.831/hour

c. OCR font typewriter

Purchase price	500.00	
40-month amortization	12.50/month	
Hourly cost (based on 176 hours per month adjusted for 75 percent utilization factor)		.095/hour

d. OCR scanner

The rental price of an OCR scanner capable of the performance in C5 is approximately $16,000 per month. The capacity of the scanner is about 600 documents/hour. The optimum size for a document for one manufacturer's device is 8-1/2" x 14". A sheet of paper of this size can accommodate 37 double spaced lines of 75 characters each. The number of records that can be typed on a sheet is a function of the number of characters in the record. The number of characters in the record is a function of the type of pre-editing the record has received. Since it was assumed that all 75 character positions in each line would be used, three blank lines were added to allow space for corrections made during input. The following calculations were made:

(1) Unedited records (325 characters/record):
5 lines + 3 = 8 lines/record or 4 records/page
or 2,400 records/hour.

(2) Partially edited records (412 characters/record):
6 lines + 3 = 9 lines/record or 4 records/page
or 2,400 records/hour[6].

(3) Fully edited records (500 characters/record):
7 lines + 3 = 10 lines/record or 3 records/page
or 1,800 records/hour.

In view of the relatively low volume of input, it would not be economical to rent an OCR scanner. Therefore, a service bureau rental of $200 per hour was used to compute the cost of this device on a per-record basis for each type of record:

Unedited records	$.083/record
Partially edited records	.083/record
Fully edited records	.111/record

e. On-line typewriter

The hardware/software configuration described in appendix H with multiprogramming capability would require at least 128K bytes of core storage. Assuming the memory capacity for servicing 20 on-line typewriter terminals plus the monitor system necessary for time-sharing, another 128K bytes of core storage would be required. This latter 128K storage plus a selector channel, storage protect, and a 2311-type disk would be dedicated to the on-line system and must be prorated across the number of terminals. Costs for these devices have been estimated as follows:

---

[6] Since, ordinarily, only complete records would be allowed on a page, the difference between 8 lines/record and 9 lines/record disappears in this computation.

128K memory module		$6,590/month
Selector channel		360/month
Storage protect		155/month
Disk		590/month
	Total	$7,695/month for 20 on-line terminals
Cost prorated by terminal		$385/month
On-line typewriter terminal		82/month
Timing adapter		23/month
Line adapter		3/month
	Total	$493/month
Hourly cost (based on 176 hours per month adjusted for 75 percent utilization factor)		$3.73/hour

3. Cost Per Record

The cost per record for an input device is calculated by dividing the cost of the equipment per hour by the hourly production rate.

a. An unedited record has 325 characters. At 6,600 strokes per hour, an operator will produce 20.3 records per hour.

Cost/record for OCR typewriter	$\frac{\$.095}{20.3} =$	$\$.005$
Cost/record for OCR		.083
	Total	.088
Cost/record for tape inscriber	$\frac{.831}{20.3} =$	.041
Cost/record for on-line typewriter	$\frac{3.73}{20.3} =$	.184
Cost/record for direct-read OCR		.167

b. A partially edited record has 412 characters. At 6,600 strokes per hour, an operator will produce 16.0 records per hour.

Cost/record for OCR typewriter	$\frac{.095}{16.0} =$	$\$.006$
Cost/record for OCR		.083
	Total	.089
Cost/record for tape inscriber	$\frac{.831}{16.0} =$	.051
Cost/record for on-line typewriter	$\frac{3.73}{16.0} =$	.233

c. A fully edited record has 500 characters. At 6,600 strokes per hour, an operator will produce 13.2 records per hour.

Cost/record for OCR typewriter	$\frac{.095}{13.2} =$	$\$.007$
Cost/record for OCR		.111
	Total	.118
Cost/record for tape inscriber	$\frac{.831}{13.2} =$	.063
Cost/record for on-line typewriter	$\frac{3.73}{13.2} =$	.283

4. Summary of Input Costs Per Record

The 20 conversion methods were analyzed for the cost of the input devices and the product of each method. The cost per record for each type of input device by major division (A-J) may be summarized as follows:

	Method and device	Cost per record
A	Direct-read OCR	$.167
B	Unedited; tape inscriber	.041
C	Unedited; OCR font typewriter plus OCR	.088
D	Unedited; on-line typewriter	.184
E	Partially edited; tape inscriber	.051
F	Partially edited; OCR font typewriter plus OCR	.089
G	Partially edited; on-line typewriter	.233
H	Fully edited; tape inscriber	.063
I	Fully edited; OCR font typewriter plus OCR	.118
J	Fully edited; on-line typewriter	.283

All conversion methods using the OCR font typewriter plus an OCR and the on-line typewriter had a higher unit cost. Therefore, C, D, F, G, I, and J were eliminated from any further consideration in the main body of the report. They are included in table I.2 of appendix I where man-machine costs are given for all 20 technical alternatives.

The remaining eight technical alternatives provide the means of making a comparison among the costs of the following basic methods:

Technical Alternative	Input Device	Form of Pre-Editing
A2 and 3	Direct-read OCR	None
B2 and 3	Magnetic tape inscriber	None
E2 and 3	Magnetic tape inscriber	Partial
H1 and 4	Magnetic tape inscriber	Full

E. Format Recognition

All major divisions A-J that have the associated secondary division 2 or 3 require processing by a format recognition program. In some instances the program would operate on partially edited records; in others, it would process unedited records.

The estimates made in appendix H for processing times for various alternatives were based on MARC II experience operating on fully edited records. The present programs at the Library of Congress (Pre-edit, Format Edit, and Content Edit) that process MARC II records use approximately three seconds/record for these functions. The processing of a partially edited record by a format recognition program adds some complexity to the present MARC II system but also duplicates part of the functions performed. Therefore, it was judged that the same amount of machine time (three seconds) would be required to process partially edited records as required to process fully edited records. The format recognition program for unedited records will be more complex than the program for partially edited records. An exact measure of how much more complex cannot be made without designing, writing, and timing both programs.

An approximation of complexity equated to machine running time was made and four seconds was allocated to format recognition processing applied to unedited records.

Therefore, an additional unit cost per record must be added to those technical alternatives that process unedited records. With a machine configuration having a rental cost of $30,000 for 176 hours of prime time, the cost per hour of the configuration equals $170. Assuming that format recognition takes an additional second of machine time to process an unedited record as compared to a partially edited record, the cost is $.047 per record.

It should be stressed that these time estimates are based on LC experience on a 360/40 DOS system not operating in a multiprogramming environment. They are subject to adjustment by more exacting timing estimates as well as variation in the equipment.

F. Sorting and Printing Costs

1. Sorting

All technical alternatives described in this chapter require sorting records by LC card number and a printout for proofing. Since the sort by card number applies across the board, the cost of this sorting has been absorbed in the cost of the hardware configuration. Any technical alternative that involves catalog comparison would also require that the records be sorted by 10 characters of the main entry to facilitate comparison of the record input from the LC card set (LC card number sequence with year) with the Official Catalog (alphabetic sequence).

Assuming only the new records per day would be sorted alphabetically and there were 2,000 such records per day, the sort time would be six minutes (see appendix H). If the machine configuration has a rental cost of $30,000 for 176 hours of prime time, the cost per hour of the machine configuration equals $170. Therefore, the cost of sorting 2,000 records in six minutes equals $17 or $.009 per record.

2. Printing

Printing costs have been calculated on a per-record basis for all technical alternatives assuming printing would be performed in a time-shared environment. The number of lines printed per minute influences the cost of printing. The estimate for this report was based upon experience at the Library of Congress in printing proof sheets (diagnostics in a format designed especially for proofing). The average speed has been approximately 420 lines per minute. Assuming that 2,000 records were in the system, a total of 48,000 lines would be printed each day[7]. At a rental price of $30,000 per month for the machine configuration for 176 hours of prime time, the cost per hour of the machine configuration equals $170. At 420 lines per minute for 48,000 lines, the print time

---

7. This is based on estimates of 24 lines per record (12 character string lines and 12 white lines; i.e., double spaced). A change in this figure will change the cost per record but not cost per line.

would be approximately 114 minutes or 1.9 hours. Therefore, the cost of printing 2,000 records would be $323 (1.9 hours x $170/hour) or $.162 per record.

The machine configuration described in appendix H assumes a multiprogramming environment. Therefore, the cost of the machine configuration would be shared between the printing operations and some other processing being performed simultaneously. It is impossible to predict what program might be running during print time, and to distribute costs between the print operation and the running program. Therefore, the cost of printing assumes a figure of an hourly cost of $30. The cost of printing 2,000 records, therefore, would be $57 (1.9 hours x $30/hour) or $.029 per record.

The costs of the technical alternatives that require comparison with the Official Catalog must be adjusted to show an additional printing cost. The printout of the records sorted alphabetically would be used as a medium for recording changes on records that have been made only in the Official Catalog record. In addition, each printed record would have to be compared with the hard copy produced after microfilming the record set to proof the machine-readable record. This comparison must be made in LC card number order, the sequence of the source data from the record set file.

A special printing technique could be used to reduce manual effort in matching these two files. It is possible to print two records side by side on computer paper 13-1/2 inches wide. In this format a printed record would require approximately 10 percent more lines than are needed for printing a single record. The tape containing the daily input of records in LC card number order would be re-sorted into alphabetic sequence on main entry resulting in two tapes, one in LC card number order and the other roughly in main entry order.

Half of the print buffer would be loaded with characters from a record from the alpha tape and the other half with characters from a record from the LC card number tape. The resulting printout would have the identical number of records printed two up in both sorts. The listing would be cut in half. The alphabetic listing would be used for the Official Catalog comparison and since it is expected that on the average only 20 percent of the records will require change, approximately 400 out of 2,000 will be modified. Each record in the alphabetic listing has an associated LC catalog card number and the changed records would be used to replace the identical record in the LC catalog card number printout. The LC catalog card number printout would then be used for proofing the source data which is also in LC card number order.

The print time and cost computed above must be modified for the two-up print. The assumption of 10 percent more lines per record printed raises the estimate of the total number of lines printed per record to 27 lines. Therefore, 2,000 new records per day would result in 54,000 lines printed each day. At a rental price of $30,000 per month for the machine configuration for 176 hours of print time, the cost per hour of the machine configuration equals $170. At 420 lines per minute for 54,000 lines, the print time will be approximately 128.5 minutes of 2.14 hours.

Therefore, the cost of printing 2,000 records two-up will be $364 (2.14 hours x $170/hour) or $.182 per record.

The cost of printing 2,000 records in a time-shared environment assuming an hourly machine cost of $30 will be $64 (2.14 hours x $30/hour) or $.032 per record.

G. Computer Configuration Costs

1. Introduction

The detailed requirements for a computer system large enough to process and hold a large centralized bibliographic data store are described in appendix H. The costs of this system are considered here. Ideally, if the configuration and the cost were a linear function of file size or processing volume, the system could start small and grow as the numbers of records processed required. In practical terms, however, the final size of the file must be considered and a system capable of expanding to that size must be predicted at the start.

2. Influence of Storage Capacity

The analysis assumes a data store that will ultimately hold a collection of one million to five million records. These records are assumed to average 500 characters in length, and they require, in addition, overhead storage for directories to locate the records. This overhead data will occupy a minimum of 10 to 15 percent of the main file.

As shown below, the approximate cost of a system capable of storing and operating on a store of one million records is approximately

$45,000 per month for one shift; or considering the storage function only, about $.045 per record per month. These estimates do not include data preparation equipment.

To store half the number of records, with a degradation in system performance because of reduced access speed, the Bryant disk alone could be used for both record storage and record location information (directories) eliminating the faster disk pack. This would reduce the cost only by 10 or 12 percent.

Cost and Configuration for One Million Records

Device	Description	Cost
Computer:	Medium scale machine, such as SDS Sigma 7, IBM 360/50, or RCA 70/45 with six tape units, card reader, card punch, and line printer	$30,000/month
Main random-access mass storage:	Large scale disk file, such as Bryant 4000 series. Capacity, 400 million bytes; estimated storage of 750,000 records.	8,350/month
Secondary random-access mass storage:	Disk pack system, such as IBM 2314, with eight drives. Capacity, 200 million bytes; estimated storage of 250,000 records plus locating information directories for one million records	5,570/month
	Total	$43,920/month
	Budgetary estimates (rounded):	45,000/month (one shift)
		~~63,000~~/month (two shifts) 49,000
		~~81,000~~/month (three shifts) 53,000

Differential Cost for an Additional Million Records

Second large scale disk file	$8,350/month
Second disk pack system	5,570/month
Total	$13,920/month
Budgetary estimates (rounded):	15,000/month (one shift)
16,000	~~21~~,000/month (two shifts)
17,000	27,000/month (three shifts)

By contrast, the capacity can be increased in increments of one million records for about $15,000 per month. For the first increment this is a 100-percent increase in capacity for a 30-percent increase in cost.

3. Influence of Processing Rates

The previous remarks have considered only storage considerations. Processing rates are obviously an additional determinant. The system described here is capable of processing about 1,000 converted records per shift in addition to performing its storage-oriented services. Hence, approximately 5,000 records per day or about 1.3 million records per year would be an upper limit for a practical system. For convenience, this figure has been rounded to one million.

The daily and weekly allocation of processing time by principal system activities is tabulated in appendix H. The principal activity is record conversion and file building. This requires most of the total time scheduled. Activities related to the distribution service occupy only a few hours per week. A token number of on-demand requests (2,000 per day) is included for testing purposes. About two hours and 10 minutes are required to process these requests.

A three-shift operation provides two considerable advantages: it minimizes the average record processing cost as well as the total elapsed time required to process a given number of records. Three shifts would permit approximately one million records to be processed in 12 months, and the computer system would cost about ~~$81,000~~ $53,000 per month.

To increase the processing capacity, it would be possible to add an additional central processor to some configurations to achieve a true multiprocessing capability. Alternatively, a second basic system complete with peripherals could be added to share the same mass memory. Since a great many of the operations performed are tape operations, even a multiprocessing approach would require additional peripherals. Therefore, it would be necessary, in essence, to duplicate the basic computer system ($30,000 per month) to achieve increased processing capacity. In this case, processing capacity would be almost doubled. True doubling would not be achieved because of increased demands on shared mass storage and the consequent increase in service times. An advantage of the dual computer approach, however, would be increased reliability; that is, some processing capability would remain even with one system down.

4. Basic Cost Schedule: Three-Shift Operation

If it is assumed that one processor would run three shifts, and a new disk system would be added every 12 months as an additional million records were processed, then four million records would be completed in just over four years. The monthly cost would start at ~~$81,000~~ [illegible] per month during the first 12 months, and be raised by $27,000 $17,000 increments every 12 months, reaching a cost of ~~$162,000~~ $104,000 per month at the end of four years. At this time the system would have four sets of disks, which would be sufficient for the four million records it would hold.

5. Basic Cost Schedule: Two-Shift Operation

In the section on organization, staffing estimates are made for a production of 10,000 records per week. This production rate could be serviced by a two-shift operation on one computer, and would have the advantages of leaving scheduled time for preventive maintenance and having slack time to make up for unscheduled down time.

In this case, the monthly cost would start at ~~$63,000~~ $49,000 per month and would be raised by $21,000 $16,000 increments every two years (about 100 weeks) reaching a cost of $126,000 $97,000 per month at the end of six years. At this time, this system would have four sets of disks, which would suffice to the end of the eighth year when it would have reached a capacity of four million records.

The computer system for the two-shift approach would cost a total of about ~~$9.1~~ $7.0 million dollars over eight years. For the same four million record final capacity, the three-shift approach would cost a total of about ~~$5.8~~ $3.8 million over four years.

6. Additional Costs

Certain one-time cost factors in computer operations depend upon the site selected. Because of wide variations in the age and utility capacity of buildings, it is impossible to assess these costs until a site has been selected. Among the factors to be reckoned with are (1) adequacy of loading docks, hallways, and freight elevators to accommodate heavy equipment, (2) electric power, (3) communication facilities, (4) floor loading capacity, (5) availability of general air conditioning, (6) ease of installing reserve air conditioning for "hot spots," (7) ceiling height, (8) room for expansion, and (9) freight and rigging costs for installation. Inadequacy in any of these conditions would result in substantial expenditures for site upgrading or the expense of moving to a different site when more space is needed.

H. System Design and Programming Costs

The costs of systems design and programming for a RECON service (assuming contractual support at $35,000 per man-year) are as follows:

Task	Man-years	Cost (rounded)
System design of procedures, hardware, software	2	$70,000
Implementation of software	14.25	499,000
Subtotal	16.25	$569,000
Software for direct-read OCR (if feasible)	3	$105,000
Total	19.25	$674,000

Chapter 6

TECHNICAL ALTERNATIVES: MANPOWER CONSIDERATIONS

A. Introduction

The alternative means of converting cataloging data to machine-readable form were discussed in chapter 5 from the standpoint of machine requirements and their costs. This chapter considers the functions requiring manpower, the staff complements needed to achieve a specified level of productivity by the major alternatives, and the unit costs for manpower in each case.

B. Functional Requirements

1. Selection of the Data for Conversion

All of the conversion alternatives would require sorting the LC record set to identify the records to be converted. As has been noted in chapter 5, the record set is arranged by card series and grouped by year within each series. Only a few of the series (notably the C, J, and K for Oriental materials) are linguistically homogeneous; all of the others are mixed. Thus it would be necessary to go through them, card by card, to group them by the languages that might be converted. Since this manual sort would be time-consuming under the best circumstances, it seems desirable to divide the record set into all of the groups that might ever be converted even though the immediate objectives of the conversion project might be quite limited.

After the cards were microfilmed, the record set would have to be reconstituted in its original order. This step would be facilitated by the fact that, since the sequence of LC card numbers would not have to be disturbed by the original sort, many of the cards would remain in sequential blocks.

Although other methods of selecting the data were considered, it could not be demonstrated that they would offer significant cost savings. Therefore, since only manual selection is applicable to all technical alternatives, it was used to determine the cost of this function. In an ongoing operation, however, the method of selection (like other phases of the process) should be reviewed constantly to insure the most efficient procedure.

2. Editing

In this analysis, the editing process comprises (1) all forms of pre-editing (that is, full or partial coding prior to input), (2) proofing for error correction, (3) post-editing to correct and augment the output of the format recognition programs, when used, and (4) editing of new data obtained as a result of comparing the interim records against the LC Official Catalog. The human effort for editing would vary with the technical alternative but a major conclusion of this study is that this function would require the largest proportion of staff in every case. The time apparently saved by raw input (direct-read OCR or keying an unedited record) followed by format recognition would largely be offset by a marked increase in the time spent in proofing and post-editing to bring the record to an acceptable level of content differentiation.

The calculations of staff requirements were based on MARC experience adjusted (where appropriate) to take account of the effects of different technical alternatives. It was assumed that full pre-editing according to the present practice of the MARC Distribution Service would require more staff than any other conversion method. This number was taken as the base staff complement; it is identical with the staff required for alternative H1. It must be stressed that experience with MARC II editing has been too brief to produce definitive figures for production rates. The estimates for the base complement were the best that could be made at the time of the RECON study.

When other editing methods were considered, assumptions were made as to the proportion of the base staff complement that would be needed to perform the function under the specified conditions.

In the absence of any pre-editing, it was assumed that the effort of proofing and post-editing would require 75 percent of the effort of full editing. This is because, without cues, a format recognition program would fail to identify data fields correctly in a high proportion of the cases. The resulting machine-readable record would be so flawed that proofing would be slow and, itself, susceptible to error because of the fatigue factor. As has already been noted, the ideal combination seems to be partial editing and format recognition processing in such proportions as to make best use of the capabilities of man and the computer.

It was assumed that the effort of partial editing would be roughly equivalent to the level of editing required in MARC I. On the basis of comparison with MARC II experience, this meant that partial editing would require about 60 percent of the effort of full editing. Partial editing would offer greater benefits than the 40-percent reduction in initial workload might indicate because the simpler coding would provide fewer opportunities for the editor to make mistakes.

In addition to pre-editing, proofing, and post-editing of the original record, the editing process must take account of the need to differentiate data added as a result of catalog comparison (described fully in 4 below). Since main, added, and subject entries would be affected by this process, changes in tags, indicators, and subfield codes might be necessary or at least would have to be considered. For the purposes of calculation, it was estimated that the staff effort required to perform this function would amount to about five percent of the effort of full editing. This estimate was based on the assumption that full content differentiation should be performed entirely by human editing.

3. Input

The physical conversion of the catalog data to machine-readable form might be accomplished by various means described in the preceding chapter. As has been noted, all of these methods would require some keyboard input at the initial conversion stage as well as the correction stage. MARC II experience was again taken as a base for calculation but an adjustment was made to take account of the fact that the keying rate on a magnetic tape inscriber is higher than on the paper tape device currently being used.

The keying effort would be directly dependent on the technical approach being used because each approach would affect the length of the record being keyed and the complexity of its coding. The most complete study of file conversion[1] indicates that the keying effort is strongly affected by the degree of complexity of the material being input. Any elements in the record that vary from straight English language texts will pose problems for the keyboard operation. By analyzing these complexities in representative records, it would be possible to calculate the degree of complexity in the average record.

An analysis of representative LC catalog cards showed that even English language records are 35 percent more complex than ordinary English text. This increased complexity was largely attributable to the fact that catalog records abound in personal names that confront the operator with the necessity of uppercase shifting and keying of unfamiliar character strings. Since the calculations of keying rates were based on present MARC experience, the complexity of ordinary English language catalog records has already been taken into account.

---

[1] U. S. Air Development Center, Rome, N. Y. Research and Technology Division. Handbook for planning file conversion. Rome, 1967. (Technical report no. RADC-TR-67-168).

Extension of the conversion effort to other languages would require that allowance be made for a reduced production rate because of the greater complexity of the records. Preliminary analyses of other types of records indicated that French and German catalog records have a complexity of 55 percent and that a group of other roman alphabet language catalog records showed a complexity of 65 percent. For the purposes of broad generalization in the cost figures for this study, 45 percent was taken as the mean degree of complexity for the records being input. As already noted in chapter 5 this results in a keying rate of 6,600 characters per hour.

It should be noted that the complexity of the record, with respect to coding that must be keyed, has a direct bearing on the probable error rate of the keying. Text that more nearly resembles straight alphabetical text should result in a lower error rate for keying than would a fully coded MARC II record. Recognition of this fact was another argument in favor of attempting to devise a method of input that would reduce the amount of pre-coding needed to achieve the final machine-readable record.

Input includes making corrections and additions to the machine-readable record, as well as keying the original record. On the basis of present MARC experience it appears that approximately one-third of the total input effort would be devoted to keying corrections to the record. It was not considered necessary, however, to calculate the requirements for these two categories separately because they require the same skill.

4. Catalog Comparison

The study of Library of Congress catalog records (appendix E) demonstrated that considerably more than half of all changes made in the Official Catalog do not result in changes in the record set. Thus, across the board, about 20 percent of the cards in the record set differ from the master records. The actual percentage of differences is directly related to the age of the records; for example, 34 percent of the 30-year-old records in the Official Catalog show changes that have not been made in the record set.

These discrepancies are especially significant when their effect is considered. The analysis showed that subject entries and added entries are most likely to be affected because, by policy, the Library does not reprint catalog records solely to show these changes.2/ Since these data elements are of prime importance for searching and retrieval, it is apparent that they should be incorporated in the machine-readable data base at the time of conversion. Failure to do so would seriously impair the quality of the records thereby imposing the task of updating them on every library that obtained cataloging information from the central bibliographic store. Although the record set can only be revised by making a record-by-

---

2. This is explained by the fact that changes in the Library's own catalogs can be made without reprinting the cards because new added and subject entries can be written at the top of unit cards without respect to the original tracings at the bottom.

record comparison with the Official Catalog, the working task force agreed unanimously that the task should be performed and its cost be absorbed as part of the basic conversion cost.

As indicated in chapter 5, parts of the record set would be converted to machine-readable form, sorted by machine on the first 10 characters of the main entries, then printed in diagnostic form. Before proofing, the records would be checked against the Official Catalog to determine what changes (if any) were required. The rough sort would reduce the effort of locating the master records by allowing the catalog editor to work in the same general area of the catalog instead of having to pursue the random alphabetical sequence of the record set.

Having located the master record, the catalog editor would be able to tell quickly in a high proportion of the cases that no change had been made on the master record. In such cases, he would simply indicate that the record had been checked and pass on to the next title. When changes were apparent, they could easily be entered as corrections on the diagnostic in most cases. In a few instances, the changes might be so extensive (e.g., lengthy additions to a contents note) that it would be more efficient to reproduce the master record by xerox or some other means to avoid tedious copying by hand.

Problems might be encountered in catalog comparison that could not be solved by the journeyman catalog editor because of language, interpretation of cataloging rules, or legibility. They would be flagged for the attention of a reviser who would clarify them in the course of checking the validity of all changes that were noted. Content designators for additions and corrections would normally be assigned in the course of the proofing and post-editing process but the task might also be done as part of the revision of catalog comparison.

If the basic conversion method did not depend on direct-reading of the record by an OCR device, it would be possible to make the catalog comparison before input to minimize the load of re-keying later.

In general, the time required to locate the master record would be relatively constant regardless of the age and language of the record involved. In some cases, however, allowance would have to be made for a slightly slower rate of locating records in difficult languages where the unfamiliar words would tend to inhibit quick finding of the record.

The difficulty in identifying, interpreting, and recording changes would be related directly to the age of the card. Older cards are more likely to appear in the Official Catalog as handwritten records, and they are more likely to exhibit pecularities in cataloging rules that make them difficult to interpret. The language of the text is also an important consideration. Foreign language records, particularly for the less common languages, would impose an additional burden on the catalog editor.

The catalog comparison effort would also run into difficulties because the Official Catalog is an active working tool. Sometimes a master record would be represented only by an out slip, indicating that it was in the process of being changed. A decision would have to be made as to whether it would be desirable to track down the record, or to allow the record to go unchanged into the data base. If provisions were made for updating records in the retrospective data base, a correction would eventually be made. The choice of action might depend on the nature of the change being made.

5. Quality Control

The final stage of the conversion process would involve a critical review of the input to insure that the machine-readable records were of a high quality. The records would already have been proofread to review the editor's work, the accuracy of input (whenever keying is involved), and the adequacy of the format recognition processing. The final review would involve verification of the foregoing steps from the standpoint of the coherence of the record. The verifier would examine the record in its own terms without direct comparison to a source document. In effect, he would ask, "Does this record make sense?" At present, in the formative period of the MARC Distribution Service, verifiers return about one out of every 10 records for some kind of correction as a result of inspecting 100 percent of the records being processed.

The high cost of inspecting every record is a strong inducement to explore all possible ways to reduce it, particularly in a large-scale project to convert retrospective records when batch processing might offer opportunities that are not available when converting current catalog records. The search for an alternative is further stimulated by the awareness that 100-percent inspection does not guarantee the detection and correction of every error. This is demonstrated by the fact that errors are discovered on LC printed cards despite repeated inspections of every record during the course of its creation.

Acceptance sampling is a well-established technique of statistical quality control. It depends on 100-percent inspection of a randomly selected lot that constitutes about 10 percent of a batch (e.g., 100 records out of 1,000). The assumption is made that the error rate in the lot is an accurate reflection of the error rate in the batch.

By defining the percentage of erroneous records that can be tolerated in the lot, it is possible to determine whether or not the entire batch should be accepted or rejected. When the percentage of error in the lot falls within the acceptable limit, the errors actually discovered are corrected and the batch is passed into the data base. When the error rate in the lot exceeds the acceptable percentage, the entire batch is subjected to 100-percent inspection to detect and correct errors.

Although the determination of an acceptable level of quality is anything but easy, the cost benefits of statistical quality control would be ample recompense for the agony of decision. The study of changes in LC catalog records suggests that approximately four percent of the manually produced records contain errors despite repeated 100-percent inspections. If this error rate can be tolerated (as, in effect, it is), it might be taken as the limit for statistical quality control.

In seeking to apply statistical quality control to the conversion of catalog records, some account must be taken of the relative importance of various types of error. An error in an access point such as main, added, or subject entry is more significant than an error elsewhere in the record. Some critical errors might be detected by machine (e.g., the check digit method of detecting an erroneous card number) but most of them would have to be found by human inspection. In any form of quality control it would be essential for the verifier to pay particular attention to the key elements of the record. In acceptance sampling it might be possible to devise a means of weighting errors to take their relative importance into account when determining the acceptability of a lot.

For the purposes of estimating the cost of quality control, it was assumed that lot sampling of 10 percent of all converted records would result in acceptance of 55 percent of the batches. This would mean that the overall quality control effort would amount to inspecting about 50 percent of the total number of records (the 10-percent sample plus total inspection of 45 percent of the remaining 90 percent equals 50.5 percent). If this reduction in effort could be achieved, the cost of the conversion project would be materially reduced.

C. Administrative Organization

1. Basic Assumptions

To calculate unit costs for manpower in the conversion effort, it was decided to create a basic staff complement capable of processing 10,000 records a week for each technical alternative. This hypothetical organization can serve as a module for determining the level of staffing needed to convert any given number of records in a specified time span.

Sections were planned to perform each of the functions of conversion. Each section was staffed at the level required to maintain the 10,000-records-a-week conversion rate, 52 weeks a year. When allowance was made for the average time taken for vacation, sick leave, and holidays by Federal employees, it was calculated that the average number of working days during the year was 223. It was judged also that one could not realistically expect peak production rates to be maintained through a working day. Six hours was taken as the period of effective daily production to allow for training, rest periods, problem resolution, fatigue, and irregularities in work flow. Therefore, a total of 1,338 effective hours per year was used to calculate production rates and unit costs.

2. Categories of Staff

The following assumptions were made about the categories of staff required to conduct a project of this nature:

a. Project Direction

It was assumed that the same level of project direction would be required regardless of the technical alternative. This office would be responsible for maintaining overall surveillance of the project, seeing that production goals were met, and resolving administrative problems. It was assumed that the following positions and grade levels would be appropriate to the responsibilities of the office:

Job title	Approximate annual salary	Federal grade	Number
Project head	$17,511	GS-14	1
Assistant project head	14,889	GS-13	1
Secretary	6,532	GS- 6	1
Clerk	4,753	GS- 3	1

b. Editing Section

The staff requirements for editing in a conversion project dealing with retrospective catalog records differ from those of one dealing with current catalog records. In the latter case, the editor benefits from the fact that the cataloger has assigned mnemonic tags to many of the data fields. In a retrospective conversion project, no such benefit would be obtainable without introducing another costly step. It was assumed, therefore, that somewhat higher grades of staffing would be required for the retrospective conversion project. The section should have the following categories of staff:

Job title	Approximate annual salary	Federal grade	Number
Head	$12,580	GS-12	1
Assistant head	10,543	GS-11	1
Supervisor	8,744	GS- 9	[3]
Editor	6,955	GS-6/7	[3]
Clerk	4,753	GS- 3	1

The ratio of supervisory staff to editors should be approximately one to 10.

---

3. The number depends on the technical alternative.

c. Input Section

The following categories of staff would be required for the Input Section:

Job title	Approximate annual salary	Federal grade	Number
Head	$8,744	GS-9	1
Assistant head	7,214	GS-7	[3]
Typist	5,316	GS-3/4/5	[3]
Clerk	4,753	GS-3	1

The ratio of supervisory staff to typists should be approximately one to 10. Therefore, the position of assistant head would not be required in alternatives A2 and **A**3 which have 9 and 10 typists respectively.

d. Catalog Comparison Section

Since requirements of catalog comparison (when applicable) are not affected by the method of input, it is convenient to present the actual numbers of staff for each category:

Job title	Approximate annual salary	Federal grade	Number
Head	$12,580	GS-12	1
Assistant head	10,543	GS-11	1
Reviser	8,744	GS- 9	1
Catalog editor	6,532	GS-5/6/7	14
Clerk	4,753	GS-3	1

The high ratio of supervisors to catalog editors (approximately one to five) would be required for the proper fulfillment of the responsibilities of this section. The journeyman editors would often be unable to interpret catalog changes correctly so their work would have to be guided and reviewed by the three supervisory staff members. The three grade levels for editors would provide a promotional ladder to take account of different levels of capability acquired through experience.

e. Quality Control Section

The workload of quality control remains constant regardless of the conversion method so the following summary shows the number of staff that would be required in each category:

Job title	Approximate annual salary	Federal grade	Number
Head	$12,580	GS-12	1
Assistant head	10,543	GS-11	1
Verifier	8,744	GS- 9	12
Clerk	4,753	GS- 3	1

The grade level of the verifiers is influenced by the responsibility placed upon them. They would have to be more experienced than editors and so should be paid at a higher rate. The ratio of supervisors to verifiers should be about one to six.

3. Staff Levels

Table 6.1 shows the levels of staffing in terms of numbers of persons required to carry out the functions of each of four major technical alternatives. Appendix I has a table showing staff for all 20 conversion methods. It will be observed that variations among technical alternatives hinge on differences in the level of staffing required for editing and input. Staff for project direction, catalog comparison, and quality control remain constant regardless of the means of converting the record to machine-readable form.

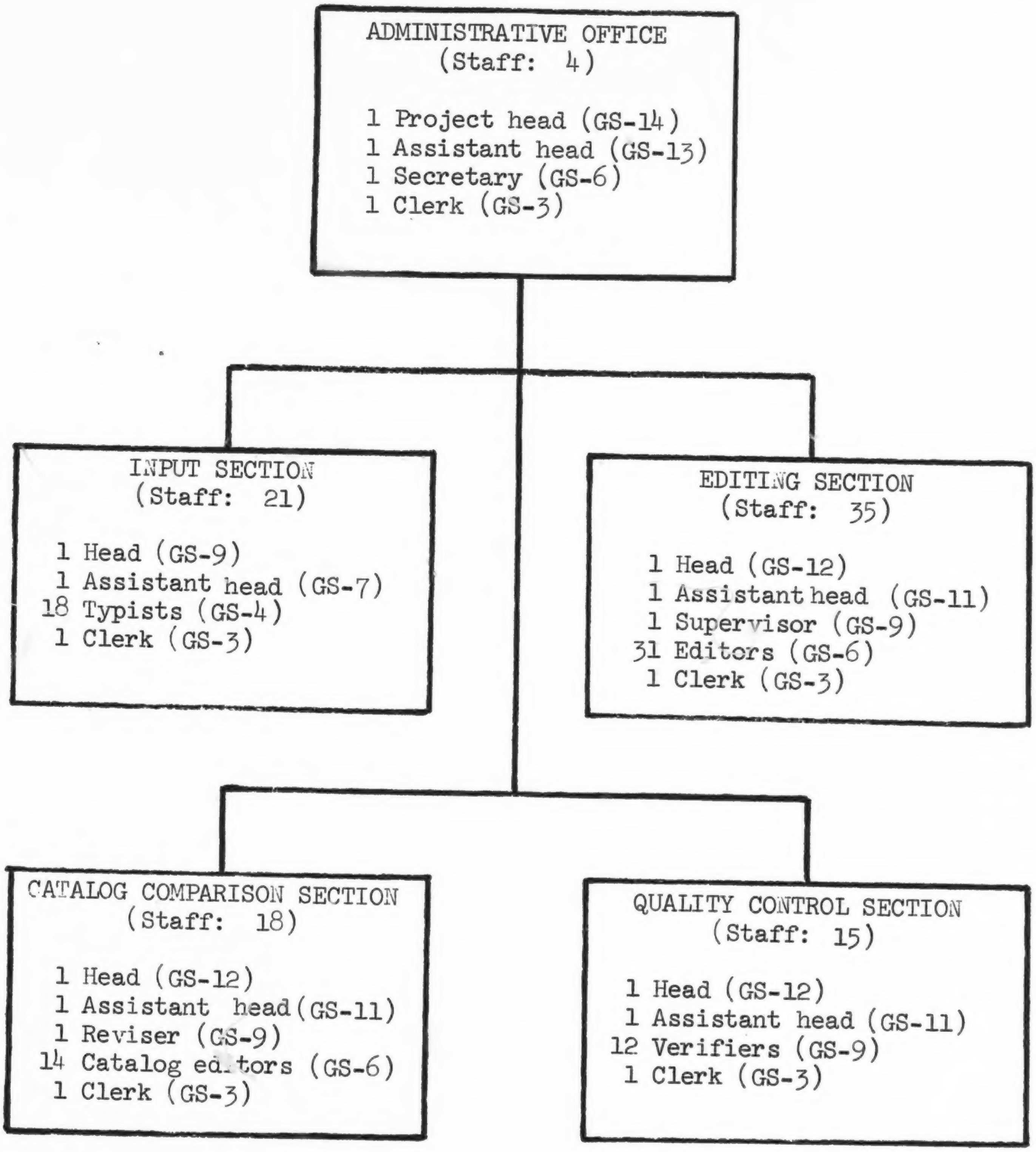
Table 6.1--Staff complements for each conversion function, by major technical alternative

Function	Technical alternative			
	A3	B3	E3	H4
Project direction	4	4	4	4
Editing	43	43	35	55
Input	12	22	21	26
Catalog comparison	18	18	18	18
Quality control	15	15	15	15
Total	92	102	93	118

It is interesting to compare the number of staff members required for each of these alternatives. As expected, alternative H4 (full editing) requires the largest staff complement. The difference between A3 (direct-read OCR) and B3 (keying an unedited record by magnetic tape inscriber) is obviously accounted for by the material difference in the number of staff members required for the Input Section. The surprising point is that the staff for alternative E3 (partial editing) is almost identical to that for A3. The significance of this similarity will be fully appreciated when machine and manpower costs are added together in chapter 7 to establish the total unit cost per record.

Figure 6.1 shows a detailed table of organization for alternative E3 which is judged to provide the optimum staff for the conversion project.

Figure 6.1--Table of organization for technical alternative E3 (Staff: 93; weekly output: 10,000 records)


ADMINISTRATIVE OFFICE
(Staff: 4)
1 Project head (GS-14)
1 Assistant head (GS-13)
1 Secretary (GS-6)
1 Clerk (GS-3)
INPUT SECTION
(Staff: 21)
1 Head (GS-9)
1 Assistant head (GS-7)
18 Typists (GS-4)
1 Clerk (GS-3)
EDITING SECTION
(Staff: 35)
1 Head (GS-12)
1 Assistant head (GS-11)
1 Supervisor (GS-9)
31 Editors (GS-6)
1 Clerk (GS-3)
CATALOG COMPARISON SECTION
(Staff: 18)
1 Head (GS-12)
1 Assistant head (GS-11)
1 Reviser (GS-9)
14 Catalog editors (GS-6)
1 Clerk (GS-3)
QUALITY CONTROL SECTION
(Staff: 15)
1 Head (GS-12)
1 Assistant head (GS-11)
12 Verifiers (GS-9)
1 Clerk (GS-3)

D. Unit Manpower Costs

1. Selection of Data to Be Converted

As was noted in section B, selection of data to be converted was assumed to be a one-time operation and, therefore, a one-time cost. It was estimated that the Card Division record set would be sorted into major language groups at a cost of $7 per 1,000 cards. After the cards were microfilmed, the record set could be reconstituted at a cost of $3 per 1,000 cards. These costs were based on the assumption that clerical staff at GS-4 or 5 level (approximately $5,000 a year) could be trained to make most of the distinctions required for sorting.

In the expectation that all of the record groups would eventually be converted to machine-readable form by some method, no attempt was made to calculate what the true cost per record would be if only part of the data base were converted. If this were done for English language records from 1960 to date, the cost of sorting all 1960 or later records to locate the desired entries would have to be prorated among the records actually selected for conversion. Since as many as eight different segments of the master data base might be converted in various combinations, it was not feasible to determine prorated unit costs for every possibility.

2. Microfilming

Although not a pure manpower cost, microfilming has been included in this section. It was assumed that microfilming would be done by a flow camera at a cost of .2 cents per record on a mass basis. This cost includes operator, machine, and processing costs. Although the working task force has some doubts about the quality of microfilming produced by a flow camera, it was decided to use this figure on the strength of a contractor's report on OCR conversion. It should be noted, however, that this cost makes no allowance for the consequences of imperfectly reproduced cards that would have to be retrieved individually from the Card Division record set so that legible copies could be made.

All of the conversion alternatives require production of a hard copy from the microfilm for proofing if not conversion. It was estimated that copies could be made on light-weight paper at a cost of one cent per record on a mass-processing basis.

3. Costs for Other Functions

Since the manpower requirements to implement each of the technical alternatives were based on a production level of 10,000 records a week, the unit costs for each function could be calculated easily. The aggregate annual salaries of all persons required for a given technical alternative were incremented by 7.5 percent for fringe benefits (a figure based on Government budgetary practices). The resulting figure was divided by 520,000 (the number of units produced in a year).

Table 6.2 shows the unit costs for all functions related to each of four major technical alternatives. A full display of unit costs for all alternatives appears in appendix I.

In determining manpower costs, the following rules were applied:

(1) where only one Federal grade level was specified, the salary of the second step of the grade was used, and (2) where two or more grade levels were involved (e.g., editors at GS-6 or 7), an average salary level was chosen.

Table 6.2--Manpower unit costs for each conversion function, by major technical alternative

Function	Technical alternative			
	A3	B3	E3	H4
Project direction	\$.090	\$.090	\$.090	\$.090
Selection[1]	.010	.010	.010	.010
Preparation[2]	.012	.012	.012	.012
Editing	.640	.640	.521	.816
Input	.138	.252	.241	.300
Catalog comparison	.265	.265	.265	.265
Quality control	.275	.275	.275	.275

1. Includes sorting the LC record set into language categories (.007) and reconstitution of the original sequence (.003).
2. Includes microfilming (.002) and making hard copy (.010).

No attempt was made to take account of increases in salaries that will inevitably occur before the project could be implemented. The figures projected in this study represent the lower limit of the cost that might actually be incurred in carrying out such a project, even in the near future. The fairly pessimistic judgments as to productivity in relation to the number of effective hours may help to minimize the effect of overall increases in manpower costs but, in actual planning, allowance would have to be made for the upward trend of salaries over the period of the project.

E. Staffing Problems

A consideration of the manpower requirements for a large-scale conversion project would be incomplete if it ignored the tremendous problems of recruiting, training, and retaining the staff necessary to carry out the project. The analysis of the staff skills for the various sections makes it abundantly clear that they are essentially the same as those required for Library of Congress cataloging and the MARC Distribution Service. Experience in both operations shows that it is extremely difficult to build and maintain high levels of staffing for operations that involve cataloging skills at almost any level. It must be recognized, therefore, that staffing a project to convert retrospective records to machine-readable form would not depend solely on the availability of funds. Regardless of where the conversion project was based, the effort to recruit staff would be handicapped by the fact that the demand for persons with cataloging skills already exceeds the supply.

F. Space Requirements

No effort has been made to estimate the costs of space, light, heat, etc., for a conversion project, but since availability of space would be a critical concern, it merits discussion.

The Library of Congress is already so cramped for space that many of its current activities have had to be relocated far from the main buildings. Even securing space for the relatively small staff of the MARC Distribution Service has been difficult. It is almost certain, therefore, that the Library could not find room within its central buildings to accommodate a conversion staff capable of processing 10,000 records a week by any of the alternatives discussed in this report.

It is not absolutely essential, however, for the staff of the conversion project to be based in the main building or the annex of the Library. Sorting and microfilming of the record set would have to be done in the Card Division in any case. These tasks could be done in off-hours. Of the other functions, only catalog comparison is dependent on being in the Library proper, so the rest of the staff could be housed elsewhere. Their isolation would entail some loss in efficiency. Editors and verifiers would be handicapped because they could not check the Official Catalog to resolve certain problems. Transporting printouts for catalog comparison between the two locations might be troublesome, and the separation of the Catalog Comparison Section and the Project Office could be a source of administrative difficulty. In the long run, it is to be hoped that the Library's space problem will be solved by the construction of a third building.

G. Conclusion

The working task force recognizes the existing demands for staff and space to maintain and even to increase current levels of cataloging as well as to expand the MARC Distribution Service. However, the benefits to be realized from retrospective conversion by the library community, including the Library of Congress itself, warrant a concerted effort to procure all the necessary resources for implementation at the earliest possible date: funds, space, and personnel.

Chapter 7

COSTS OF CONVERSION

This chapter presents a detailed summary of the combined man/machine costs per record for the four major technical alternatives and applies them to the three high-priority segments of the master data base. In addition, hardware and software costs are summarized so that the reader may find the total costs for conversion of the retrospective material in one place.

Table 7.1 gives the man/machine unit costs for input by function. When both categories of cost are combined, it appears that the unit cost of conversion by alternative E3 (partial editing plus format recognition) would be significantly lower than the other alternatives. It is also worth noting that, when all costs are considered, conversion of unedited records by magnetic tape inscriber would be slightly lower than conversion by direct-read OCR. Table 7.2 summarizes the unit costs and shows the percentages chargeable to man and machine.

To obtain the conversion costs for each of the three high-priority segments of the master data base, the estimated number of volumes in each category was multiplied by the unit cost for each major technical alternative. Table 7.3 gives the results singly and in combination.

Table 7.1--Man/machine unit costs for each function, by major technical alternative

Function	Cost per record for each alternative			
	A3	B3	E3	H4
Project direction	$ .090	$ .090	$ .090	$ .090
Selection				
Dividing record set into language groups	.007	.007	.007	.007
Remerging language groups	.003	.003	.003	.003
Preparation				
Microfilming cards	.002	.002	.002	.002
Making hard copy	.010	.010	.010	.010
Input				
Keying[1]	.138	.252	.241	.300
Machine cost of input device	.171[2]	.041	.051	.063
Editing[1]	.640	.640	.521	.816
Format recognition	.047	.047	[3]	-
Output				
Sorting	.009	.009	.009	.009
Printing	.032	.032	.032	.032
Catalog comparison	.265	.265	.265	.265
Quality control[1]	.275	.275	.275	.275
Total (rounded)	$1.69	$1.67	$1.51	$1.87

1. The unit cost takes account of additional work generated by corrections from proofing, catalog comparison, or quality control.

2. The unit cost for direct-read OCR (.167) has been increased by .004, 10 percent of the machine cost for unedited tape inscriber records (.041) because an estimated 10 percent of the records would be rejected by OCR and thus would have to be input by keying.

3. The unit cost of format recognition in E3 is too small (less than .001) to be included in this table.

Table 7.2--Manpower and machine unit costs, by major technical alternative

Category of cost	Technical alternative							
	A3		B3		E3		H4	
	Unit cost	Percent	Unit cost	Percent	Unit cost	Percent	Unit cost	Percent
Total	$1.69	100.0	$1.67	100.0	$1.51	100.0	$1.87	100.0
Manpower	1.43	84.6	1.54	92.2	1.42	94.0	1.77	94.7
Machine [1]	.26	15.4	.13	7.8	.09	6.0	.10	5.3

1. The machine and manpower costs are derived from table 7.1. The machine portion equals the sum of machine cost of input device, format recognition, sorting, and printing. All other costs in table 7.1 were assigned to the manpower portion.

Table 7.3--Total conversion cost (in thousands) for specific categories of records, by major technical alternative[1]

Conversion category	Number of records (000)	Technical alternative			
		A3 (000)	B3 (000)	E3 (000)	H4 (000)
English (1960-March 1969)	386	$ 652	$ 646	$ 581	$ 723
Romance and German (1960-June 1970)	381	644	637	574	713
English (1960-March 1969) and Romance and German (1960-June 1970)	767	1,296	1,283	1,155	1,436
English (1898-1959)	1,728	2,919	2,891	2,602	3,235
English (1898-March 1969) and Romance and German (1960-June 1970)	2,495	4,215	4,174	3,757	4,671

1. Calculated from unrounded unit costs.

Using the least expensive method (E3), English languages records from 1960 to March 1969 could be converted for an estimated $581,000. Conversion of Romance and German language records from 1960 to June 1970 would cost approximately the same amount ($574,000). The cost of converting all English language records from 1898-1959 would amount to $2,602,000.

The cost estimates of the system design and software would be constant regardless of the number of records to be converted. The cost of the software would be essentially the same for all technical alternatives except those using the direct-read OCR. For most alternatives, the total cost for system design and software would be $569,000; the cost for alternatives requiring direct-read OCR programs would be $674,000. The estimates were based on contractual support at $35,000 per man-year. An in-house effort would involve a much lower cost per man-year but would probably require a greater elapsed time because of the difficulty of recruiting and retaining programmers.

The total cost for hardware was based on the total number of records to be converted over a period of years. The price for the computer and standard peripheral equipment (tape drives, printer, etc.) would be constant but the cost of the disks would vary with the number required. The hardware cost for English language records from 1960 to March 1969 is ~~$63,000~~ $49,000 per month. Romance and German language records from 1960 to June 1967 could be accommodated in the system at no additional cost. The conversion of English language records from 1898-1959 will result in a total hardware cost of ~~$126,000~~ $81,000 per month for the aggregate data base.

These costs are based on a two-shift operation.

There is a significant similarity of the hardware/software requirements for conversion of the retrospective material and those for the LC Card Division mechanization project. The two systems differ principally in the output phase: RECON output would be records for distribution on magnetic tape; the Card Division output would use a magnetic tape record formatted for processing by a photocomposition device to produce a printed card. If a project to convert retrospective material were conducted at the Library of Congress, there is little doubt that the costs of hardware and software could be shared with the Card Division mechanization project.

Chapter 8

FUNDING AND OTHER SUPPORT CONSIDERATIONS

A fundamental recommendation of the present study is that the MARC Distribution Service for current cataloging be expanded at the earliest possible date to include data for items in languages other than English, for items in nonroman alphabets, and for nonbook materials. Although, strictly speaking, this recommendation does not affect the present task of conversion of retrospective cataloging, its implementation is extremely important in eliminating or reducing the future accumulation of cataloging data in non-machine-readable form. The cost of this expanded operation could be thought of as a regularly budgeted operation within the Library of Congress. The goal of MARC expansion is one which the Library has already accepted; the emphasis on speed in attaining this goal does not affect the financial responsibility.

The costs of actual conversion of the retrospective catalog records should be funded through the Library of Congress by appropriated funds, possibly supplemented by grant and transferred funds. Conversion of LC's retrospective cataloging data is a major aspect of the central bibliographic system currently being explored by the Library along lines first proposed in the King report[1]. Since the present study recommends the LC Official Catalog as a master data base and, further, the identification of bibliographic elements essentially according to the standard of the MARC II format, the resulting machine-readable data should be sufficient to meet the LC requirements for its central bibliographic system in terms of completeness of content and identification of bibliographic elements. The funds allocated for this purpose should be sufficient to cover input, storage, processing, updating, and maintenance of files.

The Library of Congress should not be expected, however, to support all of the costs of research and development to create the operating system required to convert, maintain, and distribute retrospective cataloging data to other institutions. The following proposals offer approaches for funding these aspects of conversion.

The library community can expect to benefit in two general ways from conversion of retrospective cataloging data. In the first place, the incorporation of this data in a machine-based central bibliographic system at the Library of Congress will, it is hoped, enable the Library to carry out its operations more rapidly. It is clear that the operations of the Library of Congress have had, for many years, a vital effect on other libraries throughout the country. In the recent past, the Library has undertaken, through such efforts as the National Program for Acquisitions and Cataloging and the MARC II system, certain functions which are primarily directed toward the national library and research community, although they may simultaneously carry actual or expected benefits for the Library itself. Because conversion of retrospective cataloging data has been studied in this report essentially from the viewpoint of its projected benefits to the Nation at large, and because this report envisions these benefits as a real possibility, funds for research and development efforts (viewed as one-time costs rather than part of the ongoing system) should be obtained from sources other than the regular budget of the Library of Congress. Examples of these costs include (1) design and programming costs for a conversion system, (2) research and experimentation on new techniques of conversion (e.g., OCR devices, format recognition) and (3) funds for a study of the problems relating to creation of a true national data store by inclusion of holdings of other libraries in the bank of retrospective cataloging data.

---

[1] King, Gilbert W., and others. Automation and the Library of Congress. Washington, Library of Congress, 1963.

Possible sources for funds to carry on this research and development work include both private and governmental agencies already active in supporting progress in the library and information science fields.

Distribution of information from the store of retrospective cataloging of data, whether this consists solely of Library of Congress information or includes holdings of additional libraries should be thought of as analogous to distribution of information through the LC Card Division. A formula based on such factors as the number of records requested, the form (machine-readable or printout) in which information is distributed, the data (LC card number, bibliographic citation, or search code) supplied by libraries in making requests, the nature of requests in terms of categories or groups (e.g., by language, date, form of material), should be devised to provide fair and reasonable reimbursement to the centralized conversion operation and data store to cover the processing of these requests. In other words, when the products of the initial conversion operation become attractive for users throughout the country and/or when the national data store concept becomes operational, a financing plan should be instituted to put these operations on a self-sustaining basis. Until this is possible, funds must be provided to enable the service to be offered to users at a nominal rather than a prorated cost.

Staffing and space are two additional support considerations that must be fully understood. Adequate staff both in quantity and quality and sufficient space in which to operate efficiently will be essential ingredients to progress in expanding the MARC Distribution Service and the conversion of retrospective cataloging.

Appendix A

DUPLICATION IN U. S. LIBRARY COLLECTIONS

This appendix summarizes various studies and reports indicating that there is a high degree of overlap among collections in libraries in the United States.

A study of patterns of duplication as they affect union catalogs published in 1942[1] shows that the number of unique titles each library contributes to a union catalog falls off rapidly as each additional library is added, and that the number of volumes in a library is positively correlated with the number of unique titles it holds. The study also shows the average percent of unique titles found in a regional catalog to be 50. The figure of 50 percent represents the relation of the unique titles in the region to the total number of titles in the libraries in the region, without regard to the duplication of titles. When the duplication has been eliminated, the percentage of unique titles rises to 75. That is, of the total number of different titles in the region, 75 percent exist in one copy

---

1. Merritt, LeRoy C. The administrative, fiscal, and quantitative aspects of the regional union catalog. In Downs, Robert B., ed. Union catalogs in the United States. Chicago, American Library Association, 1942. p. [3]-255.

only. It was calculated that the number of copies of duplicated books actually available in the several regions averages three. Thus, many titles are not duplicated at all within certain regions, but those which are duplicated may be found on the average in three different libraries in a particular region.

It was estimated that the National Union Catalog holds an average of 80.3 percent of the titles held by 11 regional union catalogs and that any given regional catalog, on the average, holds only 9.2 percent of the titles in the NUC.

Of the 11 catalogs, only Cleveland and Philadelphia catalogs were comparable in size and type of libraries included. The duplication between those two catalogs was approximately 40 percent.

In another portion of the study, Merritt shows relationships among the collections of 46 members of the Association of Research Libraries according to an "index of distinctiveness." Again, size and distinctiveness are positively correlated, or one may say that, in general, the more volumes a library holds, the more likely it is to include the holdings of other libraries, and the more likely it is to own works that other libraries have not acquired. Similarly, the smaller the library, the more likely is its collection to be duplicated in the holdings of the larger libraries, and the less likely it is to own unique titles.

A more recent study by Nugent shows that duplication among various collections is still high. The results of this study revealed "a high degree of commonality in the six [New England State university libraries'] collections,"[2] showing, for example, that a random title from one library had a 40-percent chance of being present in another randomly selected library. When current imprint samples were tested, the figure rose to 47 percent. One of the conclusions reached is that "this high degree of duplication will result in more efficient use of shared mass storage in the regional center and indicates a high return on cooperative reclassification efforts." It was projected that information about each title in the aggregate collections of the six libraries would be useful to about three of the libraries and, if only current imprints were to be processed, an average of 3.35 would be served.

Further evidence of duplication in the holdings of American libraries is provided by the experience of the National Union Catalog. In 1967, more than 50 percent of the reports to the National Union Catalog Post-1956 Imprints Section were on LC cards and subsequent searching of the remainder revealed that 82 percent (of the original 100 percent) were covered either by LC cards or reports from other libraries. Thus, only 18 percent were unique reports even at the time of reporting. By the time that a five-year cumulation of the NUC is published, fewer than 10 percent of the titles still have only a single location. The percentage of duplication of LC records would be substantially higher except for the fact that criteria for contributing to NUC reduce reporting in categories of material in which extensive duplication occurs. Similar findings have been presented in studies by Dawson[3] and Skipper[4].

---

[2] Nugent, William R. Statistics of collection overlap at the libraries of the six New England State universities. Library Resources and Technical Services, v. 12, Winter 1968, 31-36.

The facts brought out by these studies provide abundant evidence that a high degree of duplication exists in the collections of libraries of all types. It follows, therefore, that uncoordinated efforts to convert retrospective records would result in a costly duplication of effort when a multitude of machine-readable records was produced for the same bibliographic items. In addition, it is highly probable that wide variations in bibliographic description would make it difficult to identify many of these records as being for the same item (see the following page for examples of conflicting reports submitted to the National Union Catalog).

---

3. Dawson, John. The acquisitions and cataloging of research libraries: a study of the possibilities for centralized processing. Library Quarterly, v. 27, January 1957, 1-22.
4. Skipper, James. The characteristics of cataloging in research libraries. In Association of Research Libraries. Minutes of the 68th meeting, January 9, 1966, New York City. Washington, D.C., 1966. Appendix I.

ABBREVIATED SAMPLES OF VARIATIONS IN ENTRIES RECEIVED BY THE NATIONAL UNION CATALOG

NUC entry	Variations
Rao Pagdi, Setumadhava A grammar of the Gondhi language....	Rao, P. Setumadhava. Setumadhava Rao, P Madhava Rao P Setu
Ameilh, Pierre, bp., d. ca. 1401. Le voyage de Grégoire XI...[par] Pierre Ronzy....	Amelii, Petrus, patriarch of Alexandria, d. 1401? Ronzy, Pierre Petrus Amelii
The Economist (London) Oxford economic atlas of the world, prepared by the Economist Intelligence Unit and the Cartographic Dept. of the Clarendon Press....	Clarendon Press. Economist Intelligence Unit. Oxford economic atlas of the world.... Oxford University Press.
Simmons, Ward F Report on the elevated-temperature properties of stainless steels, prepared by Ward F. Simmons... "Issued under...the ASTM-ASME Joint Committee on Effect of Temperature on the Properties of Metals."	ASTM-ASME Joint Committee on Effect of Temperature on the Properties of Metals. Joint Committee on Effect of Temperature on the Properties of Metals.
Alexander de Hales, d. 1245. Alexander Minorita: Expositio in Apocalypsim....	Alexander Alemannicus (Saxo), 15th cent. Alexander von Bremen, d. 1271.
Institute on Operation and Maintenance of School Buildings, Stanford University, 1953. Institute on Operation.... Another Stanford School Planning Laboratory publication.	Stanford University. School Planning Laboratory. Stanford University. Institute on Operation and Maintenance of School Buildings. Stanford University--School of Education.
Stanford Research Institute, Stanford University. U. S. tax incentives for private foreign investment. Prepared for the Chamber of Commerce...[by A. Kenneth Beggs...	Beggs, Alexander Kenneth, 1913- Chamber of Commerce of the United States of America.
Carnegie Endowment for International Peace. European Center. Les publications officielles et la documentation internationale; travaux de la conférence de documentation réunie à Paris le 29 janvier 1951,...	Carnegie endowment for international peace. Division of intercourse and education. European center. Conférence de Documentation. Paris, 1951. International Conference on Documentation, Paris, 1951.

Appendix B

ACTUAL AND PLANNED DATA CONVERSION ACTIVITIES IN SELECTED LIBRARIES AND THEIR USE OF LIBRARY OF CONGRESS CATALOGING

A. Introduction

This is a report of a study to help in determining the desirability and feasibility of a centralized effort to convert retrospective catalog records to machine-readable form. The specific intent of this study was:

1. To characterize, both qualitatively and quantitatively, the activities of representative American libraries in the conversion of their catalog records.
2. To ascertain the qualitative and quantitative plans of libraries that contemplate catalog record conversion activities in the future.
3. To characterize the actual use of the catalog records of the Library of Congress by other libraries.
4. To determine the probable use of the machine-readable retrospective Library of Congress catalog records by libraries other than the

Library of Congress and the expected use
of the MARC Distribution Service.

Seventy libraries participated in the survey. These libraries were chosen because they were either engaged in automation activities or were believed to be actively planning for them. All types of libraries were represented: academic, public, regional processing centers, research, school, special, and state. A complete listing appears at the end of this appendix. Sixty of the interviews were by telephone; 10 were conducted at the libraries.

B. Findings

The following discussion summarizes the findings of the study. However, the results are based on such a small sample that no statistically valid inferences can be drawn from them. Generalizations can legitimately apply only to the specific libraries studied and not to the entire spectrum of American libraries. In addition, an exact tabulation of much of the data was not possible because each library was allowed to answer in its own words.

The 70 libraries answering this survey can be divided into three groups: 41 libraries were currently engaged in a conversion project; 18 libraries were planning a conversion project to begin within three years; 11 libraries had no plans to convert any records. The first two groups were questioned separately about conversion activities so that comparisons could often be made between the groups. All of the libraries were asked about their use of Library of Congress cataloging data. For any specific activity, only those libraries planning or actually pursuing this activity were asked about its execution. Thus for any particular question the actual base of responses may be quite small and the base varies often from one question to the next.

1. Types and Forms of Materials

Of the 41 libraries actually involved in conversion activities, 12 libraries said that they were converting all their records. Of the 29 libraries not converting everything, 25 libraries were concentrating on records for specific forms or types of publications, predominantly monographs and/or serials. Twenty-four were concentrating on converting imprints falling within specific time spans, almost exclusively for the period since 1960. Ten libraries were converting particular subject classes but there were no clear trends in the choice of subject classes. Thirteen of the 29 libraries were concentrating on specific languages, and of these, approximately three-quarters were concentrating on English language works. Ten libraries were using other criteria for determining what records to convert, but there were no clear trends in their choices.

Among the 18 libraries that contemplate but have not yet activated conversion programs, ten are planning to convert all of their records. Of the remaining eight libraries, six are concentrating on a specific time span; all but one in the period from 1963 to date. Five libraries plan to concentrate on monographic records. There is little interest in subject concentration, almost no interest in language concentration, and some interest in other criteria of determination. In comparing libraries with actual conversion experience with libraries in the planning stages, the latter are more ambitious about converting more records with fewer limitations.

2. Quantities of Materials

The median number of items to be processed by the libraries actually involved in conversion activities is 70,000 to 75,000; the median among the libraries that plan to convert is 350,000. The median percentage completed by those libraries that have converted is 50 percent after a median operating period of two years. With an additional median estimated time for completion of one year, the total median conversion time becomes three years. It is revealing that the median estimated period for completion among those libraries that have not yet begun their conversion activities is only one-and-a half years, despite the fact that the median estimated number of items to be processed is approximately five times the quantity being processed by the libraries that are already converting.

3. Applications

The primary stated applications for the combined groups in descending order of frequency were (1) book catalogs, (2) catalog cards, (3) facilitation of cataloging and acquisition processes, (4) information retrieval services, (5) union catalogs, (6) accession lists, (7) circulation control, (8) bibliographies, and (9) serials systems. The experienced libraries tended to favor serials systems and production of catalog cards and accession lists. The non-experienced libraries expected automation to facilitate cataloging, acquisitions, and information retrieval services.

4. Costs and Sources of Funds

The primary sources of funds for conversion among both actual and prospective converters were their own institutions, followed in order of frequency by Federal funds, State funds, and grants. Although not always clearly identified, it appears that the Federal Government, in at least some cases, is the actual or contemplated source of grants.

Questions of cost were asked only of the experienced libraries. Of 10 libraries that were able to respond to a question regarding costs for converting a single record, the range was from 48 cents to $2 per record, the average being about $1 per record. The amount of editing that preceded the input of the record seemed to create the greatest fluctuation in cost. From the wide variation in these few cost figures, it is apparent that conversion was being done in very different ways and that few (if any) of the estimates allowed for overhead or machine costs.

5. Conversion Methods in Individual Libraries

Between 55 and 60 percent of both the libraries actively converting and those planning to convert used only library personnel to plan and design their conversion project. Approximately five to 10 percent of each group relied entirely on outside personnel. The same percentages apply when considering the actual operation of the conversion project.

Approximately 35 percent of the libraries used a combination of library and outside personnel. In general, the library explained to the contractor what it wanted to accomplish and the contractor provided the technical expertise and frequently the equipment. The library was responsible for selecting and editing the records to be converted. A majority of the actively converting libraries elected to do their own keyboarding while all of those planning to convert expected to use outside keypunchers. The responsibility of programming was evenly divided between library and outside personnel. More than half (65 percent) of the libraries actively converting have established priorities for conversion; 50 percent of those planning to convert have priorities. There were no trends as far as priorities chosen except selection of current or rush materials. Approximately 65-75 percent of the libraries edited, tagged, or altered the records prior to conversion. At least half of the libraries planned to include all the catalog card elements. Approximately three-fourths said that they would include additional elements. The most frequent types of additions were (1) local control information such as location codes, accession number, copy number, or holdings and (2) bibliographical information such as notes and annotations or an indication of the language.

6. Problems Encountered

The libraries experienced in conversion were able to cite many specific technical problems. Several problems were related to input--choice between paper tape or punch cards, accuracy of the input device, conversion of both paper tape and punch cards. A second important problem

3

area was related to the very large computer storage required and computer file organization. Lesser problems were the need to standardize cataloging information from different sources, how to tag and edit catalog records, and adaptation of old computer programs. The libraries planning to convert were primarily concerned with assembling and editing catalog records, writing programs, designing a system, and acquiring data processing knowledge. There were relatively few apprehensions regarding hardware. In short the technical problems encountered by the experienced converters bear little resemblance to those expressed by the inexperienced libraries, a further indication of the need for orientation before attempting actual operations.

In regard to administrative problems, the primary problem expressed by both groups centered around personnel. The two major problems of the experienced group were lack of required specialized manpower and difficulties in coordination and communication between library staff members and the data processing specialists. Other problems mentioned were assembling the staff and planning the basic structure of the conversion project, achieving an even work flow so that the computer was used most efficiently, and convincing administrators that conversion was a good idea. Among the inexperienced libraries the two primary issues were lack of specialized personnel and the conservatism of the user. In addition, the inexperienced libraries anticipated problems in coordinating work within a network and in dealing with catalogers who dislike automation.

In addition to the common problem of not enough money, libraries experienced with conversion mentioned more specific problems such as money for additional staff or outside contractors or enough money to finish conversion quickly and economically. Several libraries were quite conservative in their plans and seemed to make no special effort to fund their conversion activities. The libraries planning to convert wanted to be sure they had adequate funds before they started any conversion project.

7. Updating and Expansion of Converted File

Almost 80 percent of the libraries actually involved in conversion activities planned to update their file. The median frequency of updating converted records was twice a month. Fifty percent planned to enlarge or refine their converted file in some other way. In decreasing order of frequency plans for expansion include (1) adding other forms of material, (2) making format changes such as building up records to MARC II, (3) expanding computer system to include other libraries in a union catalog or a network, (4) on-line terminals, (5) going backward or forward in coverage, and (6) adding [illegible]s and fragmenting the file.

8. Network Relationships

In response to questions regarding network relationships, half of the libraries in the experienced category stated that their conversion activities were related to networks or other interlibrary undertakings. Among the inexperienced libraries, two-thirds contemplated affiliation with networks. One interesting finding in regard to network relationships is that two-thirds of the experienced libraries in the university/research category are not involved in network or related interlibrary cooperative endeavors at least insofar as their conversion activities are concerned.

9. MARC Distribution Service

Fifty of the entire 70 libraries stated that they would use the MARC Distribution Service; 12 stated that they would not; and eight were not sure. The prospective subscribers planned to use the MARC tapes as a source of cataloging data for local conversion projects and the production of catalog cards and book catalogs. Diverse reasons were given for not using the distribution service: the service was thought to be too expensive for libraries with small collections; local conversion of records would be cheaper; the coverage of the service was too limited; and printed LC catalog cards could be obtained faster.

10. Use of Elements on Library of Congress Catalog Cards

Of the 70 libraries studied, 64 used Library of Congress cards or proofsheets. Only one library said it did not change any of the catalog cards received. All but three of 10 basic elements on the cards were used by more than half of the libraries, the three exceptions being the Library of Congress class numbers, the LC book or Cutter number, and the Dewey Decimal number. All of the elements, when used, were altered in a significant percentage of instances (30-60 percent). The median percentage of entries altered in some way was eight percent, although this figure ranged from less than one percent to 100 percent. Eleven of the 64 responding libraries stated they made some alterations on every catalog entry received. The most frequently changed elements were (1) series entries, (2) subject headings, (3) various features of descriptive cataloging, (4) LC class number, (5) form of main entry, and (6) choice of added author entries, in that order.

It is highly significant that 16 libraries said none of the changes they made were essential and their libraries could operate without them. On the other hand, five said all the changes they made were essential. The most frequently mentioned essential changes were classification or book number (16 libraries), form of main entry (nine), imprint (five), subject headings (four).

Regarding additions to LC cards, as opposed to changes, 43 of the 64 libraries that use LC cards stated that they add items to them, the median percentage of entries to which additions are made being five percent. The primary additions (in decreasing order) are (1) notes and annotations, (2) subject headings, (3) series tracings or added entries, (4) title entries, (5) additional copies, and (6) location symbols. Fourteen libraries said none of the additions were essential and four said they all were.

11. Prospective Use of a Service for Retrospective Records

In response to questions regarding probable use of retrospective Library of Congress catalog records in machine-readable form, 56 of the libraries stated that they would use these records, 10 said they would not, and four did not know. If a service to supply these records did not begin for two or three years, a small number of libraries said that they would not use it and an increasing number expressed doubt about using it.

In general, the projected applications were the same as those for which the libraries themselves were converting or planning to convert. The prime additional applications for converted retrospective LC records were (1) creation of data banks for network or information retrieval systems, (2) use in reclassification or recataloging, (3) use in cataloging of older materials, and (4) expansion of processing services to area libraries.

The advantages of a centralized service were thought to be elimination of the need for libraries to do their own converting and elimination of duplication of effort; reduction in cost and time of conversion; broadening of the available data base, both nationally and locally; a decrease in the need for original cataloging; simplification of reclassification; promotion of standardization.

The following problems in the operation of such a service were anticipated: (1) knowing which records had been converted and matching them to their own collection (searching time), (2) questions of systems design permitting incorporation of retrospective records at the local level, (3) costs involved in participation in general or having to buy many more titles than are in their library, (4) possible delays in the implementation of the service, (5) problems related to incompatibility of cataloging rules and practices, and (6) training, adaptation, or recruitment of operating staff. However, 18 libraries said they did not foresee any disadvantages.

Regarding the anticipated effects of the service on participating libraries, 13 said it would reduce their conversion costs and increase speed. Nine said it would have no effect because they could not wait or were already finished. Ten stated that they would hold up or eliminate their own programs or plans for conversion, depending on when the service became available. Five others believed that the service would help them to standardize their catalog record formats. Still others stated that the availability of the service would permit them to convert when it might not otherwise be feasible to do so or would allow them to catalog more with the same staff.

It was anticipated by six libraries that such a service would lead to a reorganization of their conversion project. They surmised the machine records for LC data would be obtained centrally and that local libraries might concentrate on records not covered by the service. Eighteen libraries said they would attempt to use machine-readable records with fewer changes than they make on LC printed cards.

As for priorities for retrospective conversion, the responses showed a strong correlation between the categories of records being converted or planned for conversion and what the libraries wanted the proposed RECON project to convert. In both cases, the emphasis was on English language materials (primarily monographic or serial) in reverse chronological order. There was no clear-cut preference as to subject priorities.

These views must be assessed in the light of the fact that the survey focused on the small number of libraries actually engaged in conducting or planning automation projects. Libraries that venture into this area in the next five to 10 years might have different ideas about a service to supply retrospective catalog records in machine-readable form.

C. Libraries Represented in the Survey

Information relating to the following libraries was obtained by local visits:

- Claremont Colleges
- Harvard University
- Los Angeles County Public Library
- Medical Library Center of New York
- Montgomery County (Maryland) Public Schools
- National Library of Medicine
- New York State Library
- State University of New York at Buffalo
- Tulsa City-County Library System
- Yale University

Information relating to the following libraries was obtained by telephone:

- Albuquerque Processing Center
- Argonne National Laboratory
- Bell Telephone Laboratories
- California State Library
- Cleveland Public Library
- Columbia University
- Connecticut State Library
- Cornell University
- Dartmouth College
- Enoch Pratt Free Library
- Georgia Institute of Technology
- Illinois State Library
- Indiana University
- Johns Hopkins University
- Kansas State Libraries
- Massachusetts Institute of Technology
- Michigan State University
- Nassau Library System
- National Agricultural Library
- Nevada Center for Cooperative Library Services
- New York Public Library

Ohio College Library Center
Ohio State University
Oklahoma State Library
Oregon State Library
Pennsylvania State University
Providence Public Library
Purdue University
Redstone Scientific Information Center
Rice University
Santa Clara County Free Library
Simon Fraser University
Smithsonian Institution Libraries
Stanford University
Toronto Central Public Library
U.S. Air Force Cambridge Research Laboratories
University of British Columbia
University of California, Berkeley
University of California at Los Angeles
University of Chicago
University of Colorado
University of Connecticut
University of Kansas
University of Massachusetts
University of Michigan
University of Missouri
University of New Hampshire
University of Pennsylvania
University of Pittsburgh
University of Rhode Island
University of Toronto
University of Vermont
University of Victoria
University of Washington
Upstate Medical Center Library, State University of New York
Washington State Library
Washington State University
Washington University Medical Library
Wisconsin Department of Public Instruction
Yonkers (New York) Board of Education

Appendix C

SUMMARY OF INTERVIEWS WITH CONSULTANTS

A. Introduction

The RECON Working Task Force interviewed 27 persons with experience in the field of library automation or in other fields of significance to the study. The opinions of the consultants were sought both as individuals and as representatives of particular organizations or types of organizations. The American Library Association, commercial services, research and development corporations, and a wide range of libraries were represented. Interviews were conducted in various locations with individuals or small groups by one or two members of the working task force. The list of persons interviewed is given at the end of this appendix. Their opinions are synthesized in the following pages.

B. Desirability of a Retrospective Conversion Program

An overwhelming majority of consultants favored an undertaking to convert a national data base of catalog records into a standard machine-readable format. Such a data base would:

1. Facilitate the communication and the sharing of bibliographic information by virtue of a common format.

2. Allow libraries participating in cooperative groups or networks to create a common data base conforming to recognized guidelines.
3. Facilitate retrospective acquisitions work in the same way that the MARC Distribution Service will aid current acquisitions work.
4. Provide "instant" catalogs for new libraries.
5. Provide a valuable data base for research purposes.
6. Facilitate the publication of subject bibliographies.
7. Allow libraries to post to a national data base, a procedure resulting in a true union catalog.

One consultant expressed the view that such a project should not be undertaken, at least at this time, because:

1. The problems of organizing and accessing large files have not been resolved.
2. Large categories of items, e.g., nonroman alphabet languages, cannot yet be processed.
3. Experience with MARC II should be gained before extending its use retrospectively.
4. The filing problems have not been solved.
5. Available funds should be expended on current MARC Distribution Service.

C. Centralized or Decentralized Conversion

The consultants were unanimous in recommending that conversion be done centrally. They felt that the requirements of uniformity, in both the catalog data and the machine format, and of economy of conversion dictate centralized operations.

Centralized editing for the MARC format would be a minimal requirement for uniformity, in the opinion of one consultant. All other consultants recommended centralization of the entire production operation in order to avoid duplication of software and to make the best use of manpower and equipment.

Consultants from both university and public libraries expressed the views that cooperating libraries would accept Library of Congress cataloging as a common data base. All agreed on the desirability of using the LC catalog record for the source record, realizing that there remains the problem of titles not covered by LC cataloging. Use of the LC catalog record as the source should also provide the necessary data base for the LC Card Division's automated card production project.

D. Conversion Strategy

1. Choice of Materials

Priority of printed materials over nonbook materials was assumed by the working task force and upheld by the consultants. It was further assumed that the National Serials Data Program would take care of serials and that the retrospective conversion project would concentrate on monographs.

2. Source File

As discussed elsewhere in the report, three LC files are candidates for conversion: the Official Catalog, the shelflist, and the Card Division record set.

One consultant from a major university considered the Official Catalog the only satisfactory source; another advised conversion from the record set and updating from the Official Catalog. A classified or subject approach, favored by some consultants, would obviously have to be based on the shelflist.

3. Entire File or Selected Subfiles

One consultant warned against overfragmentation of the conversion effort and expressed a preference for conversion of the entire file if it could be accomplished within a fairly short term. In general, however, consultants suggested a phased approach based on (1) language, (2) time, (3) subject, or (4) level of use. All agreed that any subset of the entire file must be readily definable (e.g., English language records back to 1960) so that users would know what records it was likely to provide.

One consultant recommended concentrating on the less common languages, e.g., Arabic, Sanskrit (in romanized form, necessarily). The remaining consultants assigned high priority to English and common roman alphabet languages.

Several consultants from university and research libraries expressed a preference for a subject approach. One felt that the subject approach would have great political and financial benefits in allowing the production of comprehensive catalogs as each subject was completed.

The essence of the recommendation of several consultants was that, if a subject approach were to be taken, different time periods should be converted for different subjects. For example, science materials become obsolete so rapidly that retrospective conversion has less value than in some other areas. One danger here is that certain classics or standard works may be missed if a time element is imposed.

Consultants from the public library area and from commercial firms, as well as some others speaking in a private capacity, expressed a preference for the conversion of high-interest modules, i.e., bibliographies or standard lists of most-used materials. Examples suggested were Books in Print (BIP) and Books for College Libraries (BCL). Such a basis might, in the consultants' opinion, make the best use of the dollars invested, since its utility for new libraries and for retrospective acquisition would indicate a large prospective market.

Some users have already encountered practical problems that lessen the utility of both BIP and BCL as selection guides for extracting a subset of bibliographic records. The former contains no Library of Congress card numbers and the numbers in the latter frequently point to LC entries that do not correspond with the book in hand. Richard Abel & Co., Inc., has been converting the 32,000 BCL records at the rate of 2,500 to 3,000 titles a week. About 50 percent of the BCL titles currently in print do not match the catalog records. This degree of mismatch is presumably due to the high percentage of titles on the list that are in the public domain and are therefore often reprinted. The Abel Company concludes that data conversion cannot be done independently of the book, at least for titles likely to be in print or frequently reprinted.

In summary, the consensus seemed to be that the most used records command the highest conversion priority. Therefore the first to be produced should be recent English language titles, with recent titles in the common roman alphabet languages next in turn. The leading exponent of the subject approach suggested a pilot project of the last five years of English language titles combined with a long-term subject approach.

E. Levels of Completeness for Converted Bibliographic Records

Two considerations evoked discussion of different levels of completeness for the converted retrospective record: (1) different levels of record identification might be attained by different conversion techniques and (2) libraries choosing to convert their own records might convert only part of a bibliographic record with the possibility that different institutions would elect different parts of the record. Obviously, such a partial record would have broader utility if it conformed to at least a minimum national standard.

One varied group of consultants defined four possible levels of the converted record:

Level 1: Full MARC editing with book in hand.

Level 2: All that can be done *without* book in hand.

Level 3: Full bibliographic data with minimal tagging (enough to allow formulating a book catalog).

Level 4: Brief bibliographic records, with sufficient tagging for circulation records, brief entry book catalogs, etc.

An illustration of the use of two of these levels is the proposed conversion of 700,000 titles by the Institute of Library Research for a five-year book catalog supplement for eight campuses of the University of California. Stage 1 of the conversion would create a level 3 record from which the catalog would be printed; stage 2 would augment the record by format recognition to a level 2 record for the permanent machine record.[1] The institute anticipates a saving of 50 percent of the cost of manual editing, even if the algorithms for automatic field recognition work imperfectly.

There was a wide spectrum of opinion on the subject of levels of record completeness. Some held that the fullest possible tagging should be accomplished by one or another means for future searching, for interchange, or as a backup for briefer records, which will be those actually used by most libraries. Others saw the brief record (level 4) as facilitating the location of items in a network and creation of brief book catalogs. One university-based consultant disapproved of establishing lower levels, while visualizing full MARC editing as a gargantuan task. He saw difficulty in enforcing the MARC II standard (level 2), if different levels were defined, and would leave development of lesser levels up to the individual library.

---

[1] See appendix G for a discussion of format recognition.

With one exception, no one advised going back to the book. One consultant suggested a cheap machine conversion plus human editing with the book for a product that would be expensive but would equal current MARC. He further commented that searching was the only real reason for conversion.

F. Local Catalog Records

One group of consultants recommended that after the LC files have been converted, the non-LC records of three or four major research libraries be converted and added to the national data base. This would presumably pick up the major portion of materials not cataloged by LC.

One consultant expressed the opinion that the National Union Catalog is not of sufficient quality to be converted without extensive editing.

The LC card number was singled out as the most useful access point or "order number" for a given bibliographic record. Where an LC card number is unknown, a search code constructed for the author, title, and other data elements could be used to retrieve the desired record. Thus, the ordering of a retrospective machine-readable bibliographic record is essentially the exact counterpart of the current system for ordering LC printed cards. However, one consultant doubted that the technology now exists for distribution of records on demand or on the basis of subscriber profiles.

A basic purpose of the library survey described in appendix B was to identify the projected use of retrospective records. These results were supplemented by comments made by the consultants who spoke for their own libraries.

Pertinent to the extent of use and/or the cost of use of the retrospective record is the degree to which the record would be accepted or would be locally changed. Several consultants were of the opinion that many libraries would accept a standard record and give up local practices. Others see their libraries continuing to change the record to conform to local practices.

G. Cost

Costs were discussed in a variety of contexts. One consultant from a commercial service saw the retrospective conversion project providing significant cost savings. Others commented on the cost of obtaining and changing the record.

One group recommended that the creation of a national machine-readable record should be funded by the government and/or foundations whether the records originated within the Library of Congress or other major libraries. The same group added that users, including commercial users, should pay only the duplication and distribution costs of the record just as users are now charged for printed cards and the MARC Distribution Service. Operators of commercial services expressed the desire to have a free hand in the exploitation of a national data base to generate a variety of products and services for sale to libraries.

H. Consultants and Their Affiliations

Richard Abel
Richard Abel and Company, Inc.
Portland, Oregon

Donald V. Black
System Development Corporation
Santa Monica, California

Ruth Blake
Tulsa City-County Library System
Tulsa, Oklahoma

Charles P. Bourne
Information General, Inc.
Palo Alto, California

Ritvars Bregzis
University of Toronto Library
Toronto, Ontario

Thomas K. Burgess
Washington State University
Computer Center
Pullman, Washington

G. R. Campbell
University of Victoria
Victoria, British Columbia

Don S. Culbertson
Information Science and
Automation Division, ALA
Chicago, Illinois

Richard De Gennaro
School of Library Science
University of Southern California
Los Angeles, California

James L. Dolby
R & D Consultants Company
Los Altos, California

George F. Farrier
Santa Clara County Free Library
San Jose, California

Paul J. Fasana
Columbia University Libraries
New York, New York

Don Gill
Los Angeles County Public Library
Los Angeles, California

Robert C. Goodwell
Los Angeles County Public Library
Los Angeles, California

Phoebe F. Hayes
Bibliographical Center for Research
Rocky Mountain Region
Denver, Colorado

Robert M. Hayes
Institute of Library Research
University of California at
Los Angeles
Los Angeles, California

Joe Hewitt
University of Colorado Libraries
Boulder, Colorado

Richard Johnson
Honnold Library
Claremont Colleges
Claremont, California

Robin McDonald
University of British Columbia
Library
Vancouver, British Columbia

Mrs. Sydney G. Marcu
New York Public Library
Branch Libraries
New York, New York

Foster M. Palmer
Harvard University
Cambridge, Massachusetts

Eugene Petriwsky
University of Colorado Libraries
Boulder, Colorado

David G. Remington
Bro-Dart, Inc.
Williamsport, Pennsylvania

Mrs. Phyllis A. Richmond
River Campus Science Library
University of Rochester Library
Rochester, New York

Frederick H. Ruecking, Jr.
Fondren Library
Rice University
Houston, Texas

Ralph M. Shoffner
Institute of Library Research
University of California
Berkeley, California

Charles H. Stevens
Project INTREX
Massachusetts Institute of Technology
Cambridge, Massachusetts

351-845 O-69—10

Appendix D

LIBRARY OF CONGRESS CATALOG RECORDS: PAST AND FUTURE

This appendix gives figures for the number of catalog records produced by the Library of Congress from 1898 through 1968 and projections of anticipated cataloging workloads from 1969 through June 1976. The data are grouped by the predominant language of the record or, in a few cases, by the type of material cataloged (e.g., music, serials).

The figures for the retrospective records were derived from LC Card Division data on the total number of cards issued annually in each card series. More than 60 different series have been issued since 1898 but the regular (unlettered) series comprises 75 percent of all cards issued since that date. Although some of the series are restricted to particular languages (e.g., C for Chinese) or types of material (e.g., Fi for films), the vast majority have no such limitation. Therefore, to arrive at the groupings shown in the following tables, it was necessary to estimate what proportion of the cards fell in each of the categories. The estimates were based on the characteristics of the special card series, analysis of several samples of the regular series, and educated guesses about the coverage of LC cataloging with respect to languages and types of material at various periods. Despite the nebulous origins of these figures, it is believed that they are sufficiently accurate for the purposes of the RECON study.

The projections of cataloging workloads through June 1976 were derived from a Processing Department estimate for fiscal 1969 which gave almost all of the required groupings. In anticipation of a steady increase in acquisitions, the figures were incremented 5 percent each year thereafter. Since the figures were rounded to the nearest thousand, however, the change from year to year is not always uniform.

Table D.1--Retrospective records and anticipated cataloging production (in thousands) by language or form, 1898-June 1976

Category	1898-1959 (000)	1960-1968 (000)	Subtotal 1898-1968 (000)	1/69-6/69 (000)	7/69-6/70 (000)	7/70-6/71 (000)	7/71-6/72 (000)	7/72-6/73 (000)	7/73-6/74 (000)	7/74-6/75 (000)	7/75-6/76 (000)	Subtotal 1/69-6/76 (000)
English	1,728	364	2,092	32	67	71	74	78	82	86	90	580
American publications	1,234	255	1,489	21	43	46	48	50	53	55	58	374
Other English publications	494	109	603	11	24	25	26	28	29	31	32	206
Roman alphabet	930	338	1,268	51	107	112	117	123	130	136	143	919
Romance and German	698	254	952	41	86	90	94	99	104	109	115	738
Other roman alphabet	232	84	316	10	21	22	23	24	26	27	28	181
Nonroman alphabet	294	250	544	27	56	58	61	64	67	70	74	477
Slavic	193	113	306	15	31	32	34	35	37	39	41	264
Other nonroman alphabet	101	137	238	12	25	26	27	29	30	31	33	213
Other forms	156	102	258	10	22	23	25	26	27	29	30	192
Music	-	38	38	3	7	7	8	8	8	9	9	59
Audiovisual	35	33	68	7	15	16	17	18	19	20	21	133
Serials	121	31	152	-	-	-	-	-	-	-	-	-
Total	3,108	1,054	4,162	120	252	264	277	291	306	321	337	2,168

Table D.2--Proposed workloads (in thousands) for the MARC Distribution Service, by language or form, April 1969-June 1976

Category	4/69-6/69 (000)	7/69-6/70 (000)	7/70-6/71 (000)	7/71-6/72 (000)	7/72-6/73 (000)	7/73-6/74 (000)	7/74-6/75 (000)	7/75-6/76 (000)	Total (000)
English	10	67	71	74	78	82	86	90	558
American publications	10	43	46	48	50	53	55	58	363
Other English publications	-	24	25	26	28	29	31	32	195
Roman alphabet	-	-	90	117	123	130	136	143	739
Romance and German	-	-	90	94	99	104	109	115	611
Other roman alphabet	-	-	-	23	24	26	27	28	128
Nonroman alphabet	-	-	-	-	35	67	70	74	246
Slavic	-	-	-	-	35	37	39	41	152
Other nonroman	-	-	-	-	-	30	31	33	94
Other forms	-	-	-	25	26	27	29	30	137
Music	-	-	-	8	8	8	9	9	42
Audiovisual	-	-	-	17	18	19	20	21	95
Total	10	67	161	216	262	306	321	337	1,680

Table D.3--Workloads (in thousands) for retrospective conversion and proposed MARC Distribution Service, by language or form

Category	1898-1959 (000)	1960 to beginning of MARC (000)	RECON subtotal (000)	MARC output through June 1970 (000)	1898-June 1976 (000)
English	1,728	386	2,114	558	2,672
American publications (April 1969)[1]	1,234	266	1,500	363	1,863
Other English publications (July 1969)[1]	494	120	614	195	809
Roman alphabet	930	518	1,448	739	2,187
Romance and German (July 1970)[1]	698	381	1,079	611	1,690
Other roman alphabet (July 1971)[1]	232	137	369	128	497
Nonroman alphabet	294	481	775	246	1,021
Slavic (July 1972)[1]	193	225	418	152	570
Other nonroman alphabet (July 1973)[1]	101	256	357	94	451
Other forms	156	157	313	137	450
Music (July 1971)[1]	-	55	55	42	97
Audiovisual (July 1971)[1]	35	71	106	95	201
Serials	121	31	152	-	152
Total	3,108	1,542	4,650	1,680	6,330

1. Proposed beginning date of MARC Distribution Service for each category is shown in parentheses.

Appendix E

CHANGES IN LIBRARY OF CONGRESS CATALOG CARDS: THEIR EXTENT, METHOD, AND TYPES[1]

A. The Problem

Catalog records are never immune from change as long as they are part of a living catalog. Regardless of their age or insignificance, they may be affected by the cataloging of other items and thus are always susceptible to alteration. In the Library of Congress, a change may result in a revised reprinting of the record, or it may be made by hand in one or more of the card catalogs. Since catalog maintenance is a heavy chore in the traditional system however it is done, it may be expected to constitute a significant workload in keeping a file of machine-readable records up to date.

The present study has a dual purpose. First, it attempts to quantify the workload of updating so that allowance can be made for the staff and machine time required to cope with it in the MARC system. Second, the study seeks to show the extent of difference between the Card Division record set and the Official Catalog for cards of various ages.

---

1. Originally prepared by the Technical Processes Research Office of the Library of Congress for internal use.

To satisfy the first requirement, the study seeks a basis for estimating what proportion of a given body of catalog records might be changed in a specified period by analyzing random samples of catalog cards produced at various intervals during the past 30 years. Although the policies governing some of these changes are no longer in effect, it is believed that the findings give a useful indication of what may be expected in the future.

The second point (the difference between the record set and the Official Catalog) has not previously been studied. Persons connected with cataloging are well aware that the two files are far from being identical but the extent of the difference has never been quantified. Since the record set (or its equivalent in the form of stock cards) has been suggested as the source for retrospective records that may be converted to machine-readable form, information about the difference is crucial to evaluating the adequacy of this approach.

B. Methodology

What mattered in this study was whether a catalog record had been changed after its initial printing. To estimate this proportion for records of various ages, five random samples were drawn from the regular card series for the years 1938, 1948, 1958, 1966, and 1967. The cards for the two most recent years were chosen because the volume of short-range updating is most relevant to immediate planning for the MARC system. The three earlier groups were chosen to permit estimation of the rate of change as catalog records age. Addendum 1 describes the considerations in selecting the regular card series for investigation, the determination of sample sizes, the degree of reliability and precision obtained, and the method of generating the samples.

After stock cards were obtained, each of the five samples was divided into three language categories: English; other roman alphabet languages; and nonroman alphabet languages. Cards were assigned to these categories on the basis of the language that predominated in the body of the entry. The results are shown in table E.1.

Table E.1--Language categories[1] in five samples of Library of Congress cards, by card series

Card series	All languages		English		Other roman alphabet languages		Nonroman alphabet languages	
	Number	Percent	Number	Percent	Number	Percent	Number	Percent
1967	381	100.0	158	41.5	166	43.5	57	15.0
1966	443	100.0	208	46.9	178	40.2	57	12.9
1958	351	100.0	155	44.1	115	32.8	81	23.1
1948	459	100.0	248	54.0	166	36.2	45	9.8
1938	523	100.0	352	67.3	166	31.7	5	1.0

1. Cards were assigned to a language category on the basis of the language predominating in the body of the entry.

The language division was primarily to guard against the possibility that differences in the composition of the samples might have an effect on the extent of change. The subsequent analysis indicated that this was not a problem and, in any event, the distribution of language categories seemed appropriate to the periods, with the possible exception of the high proportion of nonroman titles in 1958 sample.

The language groups also offered some opportunity to determine whether the proportion of change differed among languages. It should be noted, however, that the initial sample sizes are not large enough to invest the analysis of the subsamples with any great reliability.

After this preliminary analysis, the stock cards for all five samples were searched in the Official Catalog and the 1948, 1966, and 1967 samples were also searched in the shelflist. When a stock card differed from the official main entry or the shelflist contained copy information, the variant information was noted on the stock card for later analysis. Of course, in tabulating changes, a revised reprint was counted even when the stock card and the official main entry had the same information.

C. Findings

1. Extent of Change

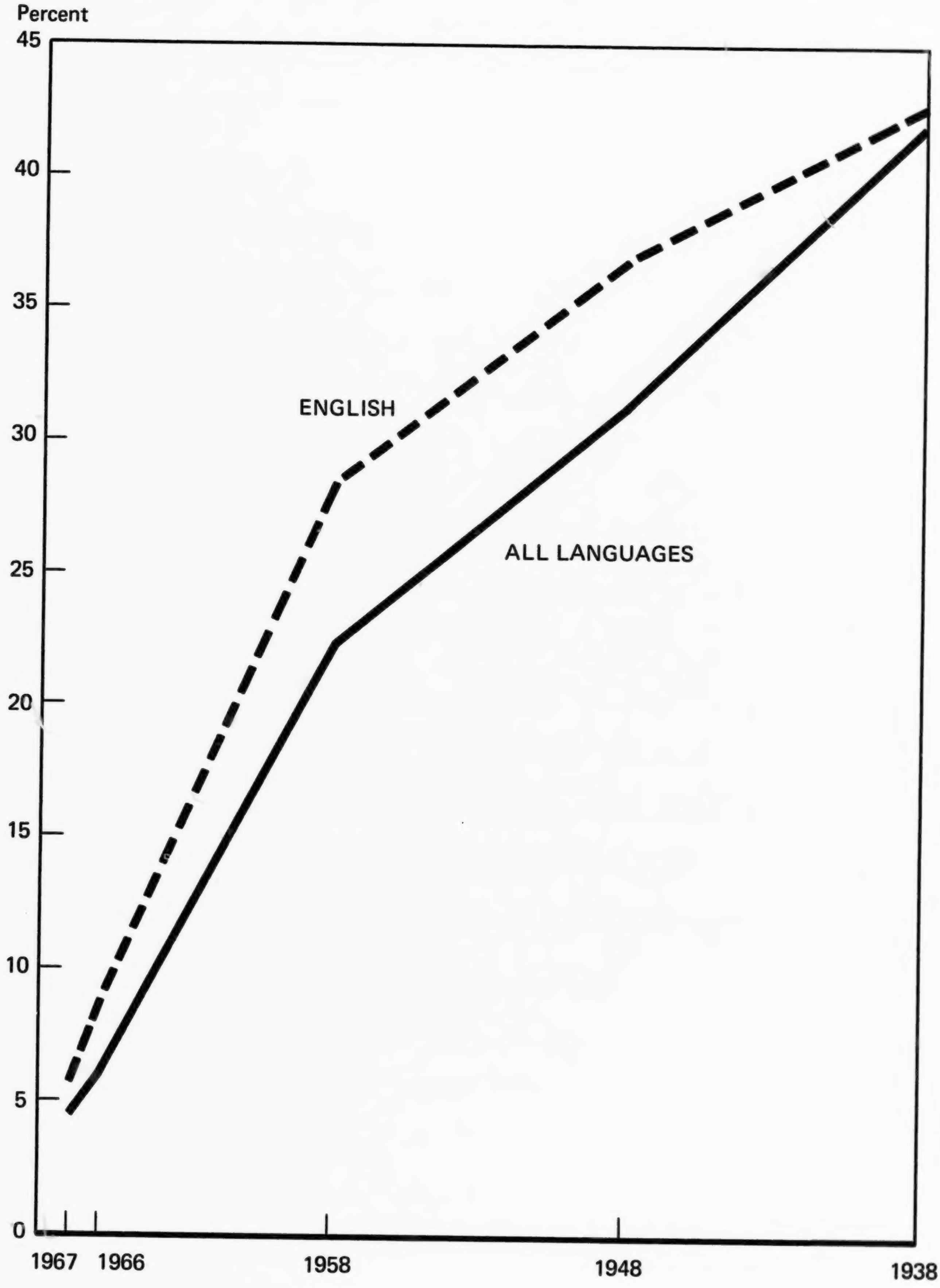
The analysis of changes affecting the groups of sample cards reveals striking evidence of both the extent of change and its persistence over long periods. A study of table E.2 suggests that the rate of change may be higher in the first years of the life of a group of catalog records, but after 10 years the rate seems to stabilize at one percent a year. Investigation of samples of older catalog records will be required before it is possible to establish at what point the trend line shown in figure E.1 tends to level off.

On the basis of this analysis, it is estimated that between 4.5±2.0 percent of catalog records put in machine-readable form will have to be updated in the first year. In view of the fact that the initial input to the MARC system will comprise only English language titles, which seem to be subject to more immediate change, it would seem prudent to

Table E.2--Extent of change in five samples of Library of Congress cards, by card series and language category

Card series and language category	Total number	Not changed		Changed	
		Number	Percent	Number	Percent
1967	381	364	95.5	17	4.5
English	158	149	94.3	9	5.7
Other roman alphabet languages	166	159	95.8	7	4.2
Nonroman alphabet languages	57	56	98.2	1	1.8
1966	443	416	93.9	27	6.1
English	208	190	91.3	18	8.7
Other roman alphabet languages	178	172	96.6	6	3.4
Nonroman alphabet languages	57	54	94.7	3	5.3
1958	351	273	77.8	78	22.2
English	155	111	71.6	44	28.4
Other roman alphabet languages	115	96	83.5	19	16.5
Nonroman alphabet languages	81	66	81.5	15	18.5
1948	459	315	68.6	144	31.4
English	248	157	63.3	91	36.7
Other roman alphabet languages	166	129	77.7	37	22.3
Nonroman alphabet languages	45	29	64.4	16	35.6
1938	523	304	58.1	219	41.9
English	352	202	57.4	150	42.6
Other roman alphabet languages	166	99	59.6	67	40.4
Nonroman alphabet languages	5	3	60.0	2	40.0

Figure E.1--Percentage of change in five samples of LC printed cards



assume that the higher figure is more accurate. It is worth noting, however, that the differences among language groups seem to be equalized in the long run.

In considering the significance of the findings on extent of change, two contradictory points should be kept in mind. On one hand, policies governing changes have been modified from time to time during the long history of Library of Congress cataloging. Thus, to the extent this is true, a study of past changes is an imperfect guide to the future. Particularly important is the fact that the application of the Anglo-American cataloging rules now makes it unnecessary to revise a corporate heading to show the latest form of name.

On the other hand, it was apparent that many of the cards had been changed on more than one occasion. No attempt was made to tally these instances because it was not always possible to determine when they occurred. It may be said, however, that the true workload of updating represented by these samples was greater than table E.2 reveals. While it cannot be asserted that these conditions offset one another, for the purposes of prediction they do have a counter-balancing effect.

2. Methods of Change

Changes in LC catalog records may result in revised reprints or they may be limited to typed or handwritten additions and corrections in the Library's own catalogs. Revised reprints are stimulated primarily by changes in main entry, title, or other elements necessary for correct identification of the book. A complete list of the criteria for revised reprints appears in Processing Department Memorandum No. 31 (see addendum 2).

The restrictions on revised reprinting have been imposed for administrative reasons; they do not constitute a judgment that other kinds of changes are unimportant. Changes in added and subject entries, contents notes, classification numbers, etc., are all essential to the integrity of the catalog records they affect, and plans to convert retrospective records to machine-readable form must take such changes into account.

Figure E.2 shows the proportion of change by each method in the five samples. The sum of the two proportions for each sample equals the percentage of change shown in table E.2. The enormous spread between the proportion of manual changes and the proportion of revised reprints in the 1938 sample apparently results from differences in policies about correcting catalog records.

In all but the latest sample, the majority of changes on catalog cards in the sample did not result in revised reprints. Thus there is no doubt that the records in the Official Catalog are significantly different from the cards in the record set.

3. Types of Change

Although no claim can be made for the statistical reliability of the data on types of change, table E.3 gives an indication of the distribution of changes with respect to the cataloging data elements affected. Note that this analysis is based on the aggregate number of changes, not the number of records changed. In tabulating these data, one change was

Figure E.2--Percentage of changes by general types in five samples of LC printed cards

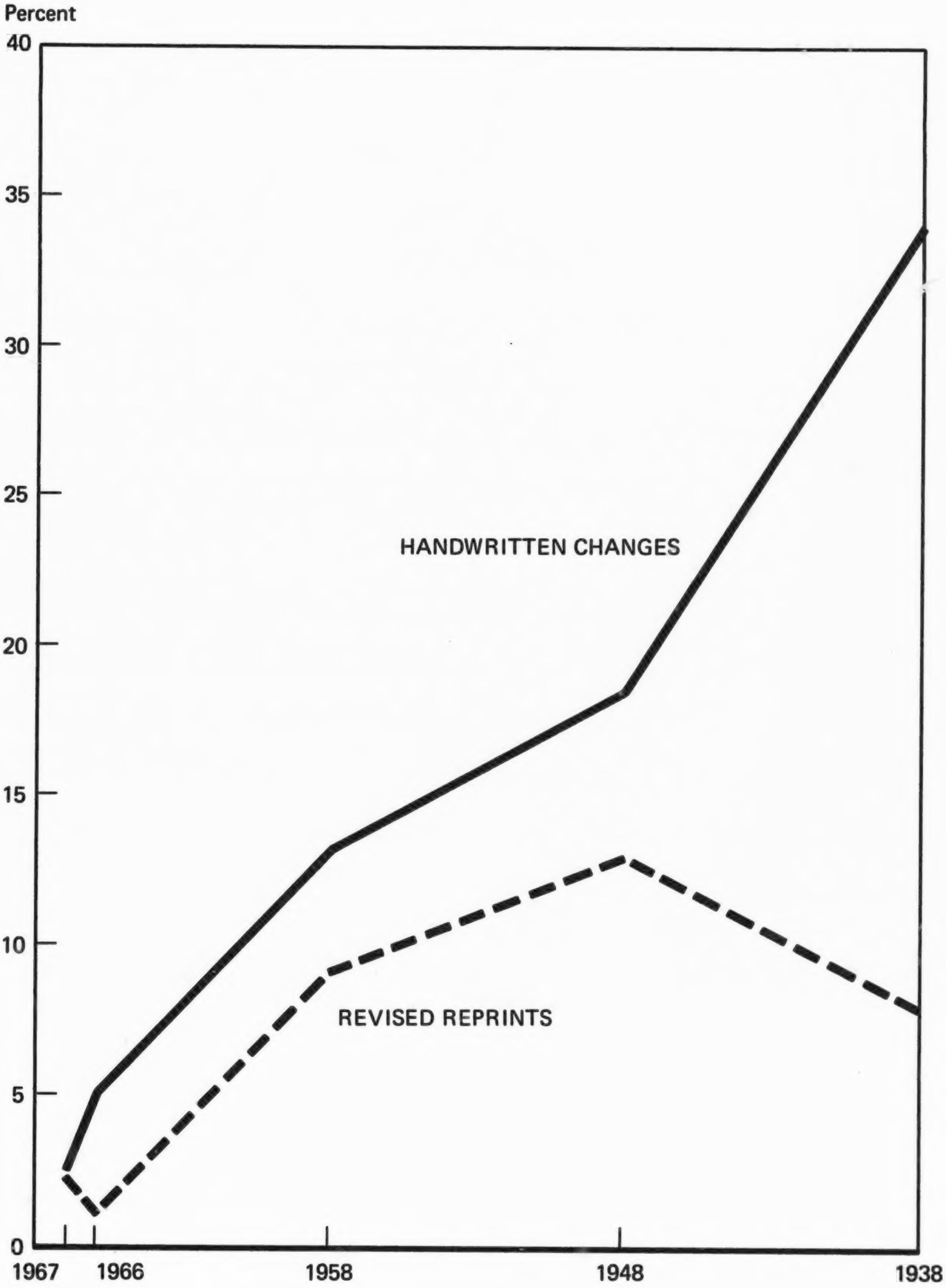


Table E.3--Data elements affected by changes in five samples of Library of Congress cards[1]

Data element	1967		1966		1958		1948		1938	
	Number	Percent	Number	Percent	Number	Percent	Number	Percent	Number	Percent
Total	21	100.0	36	100.0	117	100.0	257	100.0	382	100.0
Main entry	3	14.3	5	13.9	25	21.4	63	24.5	105	27.5
Body of entry	4	19.0	2	5.5	14	12.0	29	11.3	18	4.7
Collation	3	14.3	5	13.9	12	10.2	20	7.8	8	2.1
Series statement	-	-	2	5.5	3	2.5	4	1.5	7	1.8
Notes	3	14.3	5	13.9	12	10.2	25	9.7	5	1.3
Subject heading	1	4.8	6	16.7	34	29.0	67	26.1	145	38.0
Added entry	4	19.0	7	19.4	12	10.2	22	8.6	58	15.2
Classification number	1	4.8	1	2.8	1	0.9	11	4.3	15	3.9
Dewey number	-	-	-	-	1	0.9	2	0.8	2	0.5
Added copy[2]	1	4.8	1	2.8	1	0.9	5	1.9	4	1.1
Dash entry	-	-	-	-	1	0.9	1	0.4	5	1.3
Other[3]	1	4.7	2	5.6	1	0.9	8	3.1	10	2.6

---

1. Includes all changes that could be identified on each changed card tallied in table E.2, except for those on a few revised reprints which could not be identified.
2. Limited to copy statements on the official main entry card.
3. Includes filing title, full name note, national bibliography number, format error, and control information.

recorded for each modification on a record. For example, when the closing of the record of a multivolume set involved changes in imprint date, collation, and contents note, three changes were counted.

Although all data elements on a record are susceptible to change, the analysis shows that some are more affected than others. Changes in subject headings rank at or near the top of the list in all but the latest sample. It will be recalled that this category of change does not result in a revised reprint although such changes are made if the record is reprinted for some other reason. This fact deserves considerable weight in evaluating the adequacy of various files for retrospective conversion.

D. Implications

The findings of this study provide convincing evidence that catalog records are not immutable and that change is a fact of life in a functioning catalog. The ability to accommodate this change is an essential requirement of a viable system for storing these records in machine-readable form. Therefore, the inexorable character of change in catalog records must be taken into account in designing the organization of machine-readable data files and the means of accessing them. Only if this is done can additions, corrections, and deletions on records of any age be made quickly and efficiently.

The study also establishes the fact that the Official Catalog differs materially from the Card Division record set in the accuracy and currency of its data. Therefore, even if projects involving the conversion of retrospective catalog records begin with the record set, changes in the

Official Catalog cannot be ignored without risking a significant loss in the quality of the cataloging information, especially on older records.

Addendum 1

SAMPLING METHODOLOGY

A. Choice of Data Base

In the last 70 years the Library has issued cards in 55 different series, representing many categorizations of its catalog records. Of these, 19 were used in 1967. The most active series in 1967 includes approximately 115,000 entries; the least active, only 38. To obtain a sample representing all card series would require meticulous stratification. To avoid this exercise, it was decided to limit the samples to cards in the regular (unlettered) series. This decision to simplify the drawing of the sample seemed justified on several other grounds:

1. The regular series comprises the largest body of catalog cards (approximately 77 percent) of the total number printed since 1898.
2. Many of the other current series (C, HE, J, K, NE, and SA) are used almost exclusively for records using nonroman alphabet languages that will not be put into machine-readable form in the immediate future.
3. Still other series (e.g., A) may be assumed to have characteristics similar to those of the regular series.

B. Sample Size, Confidence, and Precision

The percentage of a total population that exhibits a specific characteristic can be estimated by analyzing a simple random sample. The size of the sample is determined by the size of the population, the anticipated percentage that will have the characteristic, and the degree of confidence and precision desired. Table E.4 shows the data for the five samples used in this study. The confidence level for all samples is 90 percent; that is, it is estimated that 90 out of 100 random samples of similar size would yield findings of the same degree of precision.

Table E.4-- Sampling table for estimating percentage of change in five series of Library of Congress cards at a confidence level of 90 percent[1]

Year of series	Number of cards[2]	Expected percent of change	Precision required	Sample size
1938	39,775	40.0	3.5	523
1948	45,811	30.0	3.5	459
1958	61,503	20.0	3.5	351
1966	99,000	7.0	2.0	443
1967	114,999	6.0	2.0	381

1. Derived from Brown, R. Gene, and Lawrence L. Vance. Sampling tables for estimating error rates or other proportions. [Berkeley, Calif.] Institute of Business and Economic Research, University of California, Berkeley [c1961].
2. Data from Card Division.

The degree of precision is ± 2.0 percent in the 1966 and 1967 samples and ± 3.5 percent in the 1938, 1948, and 1958 samples. This difference had to be accepted to keep the sample sizes within bounds. The samples of the earlier card series would have to be three times larger to obtain a precision of ± 2.0 percent. The degree of precision is absolute; that is, it is on the same scale as the estimated proportion of change.

Thus, a proportion of change expressed as 5.0 ± 2.0 percent represents a range from 3.0 to 7.0 percent.

C. Selection of the Samples

A table of random numbers[2] was used to generate the five samples for this study. By drawing each sample separately, it was possible to consider the five-digit numbers in the table as the second part of the LC card number for the series in question. A slight bias occurred in the sample for the 1967 series which includes approximately 15,000 cards with numbers larger than 99,999 (the largest number in the table).

---

[2] Rand Corporation. A million random digits with 100,000 normal deviates. Glencoe, Ill., Free Press, 1955.

Addendum 2

LIBRARY OF CONGRESS

PROCESSING DEPARTMENT

Department Memorandum No. 31

March 29, 1944
Revised
August 26, 1963
Revised
March 9, 1964

PROCEDURES FOR REPRINTING LC CARDS

The following procedures for revising and reprinting catalog cards are effective immediately. Three categories of cards to be reprinted are established: offsets, resets, and revised reprints.

Revised reprints will be prepared to replace cards already in the Library of Congress catalogs, and will also be distributed to depository libraries, the Union Catalog Division, and the Cumulative Catalog Section of the Catalog Maintenance Division for the book catalogs. Whenever any correction is made that justifies this replacement (see C2 below) the correction will result in a revised reprint even though the whole catalog entry will not normally be reviewed in depth to see whether other corrections might also be in order.

A. OFFSETS

1. Origin

Originate in the Card Division

2. Types included

Cards to be reproduced photographically without change to replenish stock. These will include cards with typographic or other errors not affecting the filing of the main or secondary entry and otherwise so minor that they can be ignored; cards required by the Subject Cataloging Division to prepare changed or corrected subject entries; and cards required by the Catalog Maintenance Division to prepare adapted sets and corrected replacements involving change of call number or other changes not calling for resetting or revised reprinting.

B. RESETS

1. Origin

   a. Originate in the Card Division to replenish stock when record card is too poor to photograph.

   b. Originate elsewhere when corrections too minor to cause the card to be treated as a revised reprint are to be made.

2. Types included

   a. Cards to be reset without change to replenish stock when there is no satisfactory card to photograph.

   b. Cards to be reset with minor changes when needed to replenish stock* and the Reprint Unit of the Card Division has been notified that corrections of the following kinds are in order:

      1. Changes in the heading that do not affect the filing, such as addition of date of death, deletion of such designations as Mrs., Sir, etc., addition or deletion of inc., etc.,

      2. change in title not affecting filing,

      3. minor change in accents, punctuation, or capitalization,

      4. change in imprint in form but not in fact,

      5. change in illustration statement in collation,

      6. change in size,

      7. minor change in running time for films or number of frames for filmstrips,

---

* Corrections of the kind described here are made on the appropriate cards in the Library of Congress catalogs by the catalogers or, at their direction, by the Catalog Maintenance Division.

8. change in series note in form but not in fact,

9. minor additions or changes in notes, including addition of title transliterated note,

10. addition of contents note,

11. addition of another issue, copy, or microfilm copy,

12. subject added or changed,

13. added entry (including series) added, changed, or deleted,

14. addition or change of LC classification number,

15. addition or change of Dewey classification number,

16. addition of dagger when a card printed from cooperative copy is adapted for LC.

3. Procedure

a. The Reprint Unit searches the Official Catalog before resetting a card if there is reason to think a change has been made.

b. The descriptive cataloger or member of the Decimal Classification Office, or Editorial Section of the Subject Cataloging Division notifies the Reprint Unit of any changes made after the date of this memorandum on any card printed in the two current series.

c. If the change is to be made in all catalogs, the descriptive or subject cataloger asks the Card Preparation Section of the Catalog Maintenance Division to correct the cards in all catalogs.

d. The Reprint Unit determines whether the correction shall be ignored until card stock is exhausted or whether stock shall be killed immediately, but will kill stock for cases 12 and 14 above when requested by the Subject Cataloging Division and 15 when requested by the Decimal Classification

Office. If resetting is delayed, the record card is stamped either "See Official Catalog before resetting," or "See attached card for corrections."

e. The Inventory Section of the Card Division prepares the card for the printer, estimating, adding symbols, etc. The symbol added to cards reset without change, e.g. [44d2] indicates the year of reprint, number of hundreds printed previously and number of hundreds printed at this printing. If any change has been made, "a" (i.e., addition) is prefixed to the symbol, e.g., [a44d2] and cards are replaced in Card Division catalogs only. A long dash in the card number is used on all resets.

C. REVISED REPRINTS

1. Origin

a. Originate in the Descriptive Cataloging Division when revisions are made.

EXCEPTION: The symbol "rev" is added to the card number when cards are reprinted for corrections before distribution to the Library's catalogs or when cards are reprinted to eliminate duplication of card numbers.

2. Types of corrections

a. Main entry changed (e.g., from corporate to personal author; author and title to title entry; or vice versa),

b. heading changed in any significant way, by correction of error in spelling or date, addition or deletion of birth date or distinguishing phrase,

c. change in title or title transliterated note if it affects filing,

d. addition or deletion of subtitle,

e. addition of author statement, editor statement, or statement of illustrations,

f. change in paging,

g. important additions or changes in notes,

h. addition of indexes and supplements,

i. entries opened or closed,

j. errors in card numbers corrected,

k. card "Printed for Card Division" adapted,

l. changes such as those listed under B2b when the corrections are important enough or numerous enough to warrant replacing all copies of the cards in the LC catalogs and including a revised entry in the book catalogs,

m. changes such as those listed under B2b when cooperatively printed cards are being adapted and the changes are numerous or difficult to incorporate.

3. Procedure

a. The descriptive cataloger notifies Reprint Unit to kill stock immediately.

b. Following descriptive revision, the card (and book, if needed by the descriptive cataloger) is forwarded to the Subject Cataloging Division, and from there the card is sent to the Reprint Unit or to the Coordinator of Cooperative Cataloging.

c. For revised reprints, "rev" is printed at the end of the card number. When cards in the Ca unrev'd series are revised they are reprinted with current card numbers, and do not indicate a previous printing.

d. Revised reprints are distributed (to Catalog Maintenance Division, Union Catalog Division, and depository catalogs) according to the distribution of new cards.

When the Card Division cannot locate the Official main card, or when it is not suitable copy for the printer, the Descriptive Cataloging Division provides a replacement, which may be reset or reprinted revised.

Cards printed in Far Eastern and Indic languages that are necessarily produced photomechanically are reset or revised in accordance with the above criteria but with special procedures involving respectively, the Far Eastern Languages Section and the South Asian Languages Section of the Descriptive Cataloging Division.

Addendum 3

COPY INFORMATION

Notations about the number and location of copies of cataloged items are largely confined to the shelflist. Copy information appears in the Official Catalog only when more than one call number or special location is involved. Although shelflist notations about copies may be made when the original record is being prepared, they are often added later and thus effect a change in the record. Since a full-scale bibliographical store for the Library should include this kind of information, it was decided to check three of the samples in the shelflist to determine how often copy information had been added after completion of the original catalog record.

In the 1966 and 1967 samples, 11.5 percent of the records (51 of 443 and 44 of 381 respectively) had been changed at least once to add copy information in the shelflist. In the 1948 sample, 15.7 percent of the records (72 of 459) had been changed in this way. These figures show only the number of records affected but the actual workload was heavier because, in a sizable number of cases, copy information had been added to the same record on more than one occasion.

The findings of this partial analysis help to quantify the additional burden of updating that will have to be assumed if the file of machine-readable records is to perform the functions of the shelflist.

Appendix F

COMPLETENESS OF MACHINE-READABLE CATALOG RECORDS

In developing plans for conversion of retrospective records, the possibility exists that not all data for bibliographic items may be recorded in machine-readable form with the degree of completeness specified by the MARC II format. Records might be created with a lesser degree of differentiation of the data (that is, simplification of the tags, indicators, and subfield codes) and/or with some limitation on the bibliographic data as might occur when a brief shelflist record is made.

Lack of bibliographic data may deprive a record of the richness of detail that would enhance its usefulness but it would not cause the same kinds of problems that would arise from variations in machine format. For example, if some records have tags that are less precise than those in other records, all records must be processed at the lowest common denominator. On the other hand, lack of a data element that may actually apply to an item (such as an index note that could only be made by going back to the book) does not preclude the processing of those records that do have a fixed field containing this information.

For discussion purposes, the working task force felt the need to define levels of encoding detail in relation to the conditions under which conversion might occur. Consideration was also given to an attempt to describe a minimal standard for conversion in local institutions.

Three levels of standards were tentatively defined as follows:

Level 1 involves the encoding of bibliographic items according to the practices followed at the Library of Congress for currently cataloged items, i.e., the MARC II format. A distinguishing feature of level 1 is the inclusion of certain content designators and data elements which, in some instances, can be specified only with the physical item in hand.

Level 2 supplies the same degree of detail as in level 1 insofar as it can be ascertained through an already supplied bibliographic record. This means that in some cases the following content designators and data elements specified in MARC II cannot be supplied from existing catalog records to be converted: (1) language, (2) index, (3) subject as main entry, (4) fiction, (5) form of reproduction (e.g., large print), and (6) form of content.

Essentially, however, the remaining tags, delimiters, indicators, subfield codes and data elements could be assigned to retrospective records with no reference to the physical item.

Level 3 would be distinguished by the fact that only part of the bibliographic data in the original catalog record would be transcribed. In addition, content designators might be restricted to those tags necessary to identify the data elements in the following list:

Main entry

Short title

Edition (transcribed to the word "edition" or
its equivalent)

LC card number, if it is available

Imprint: place, publisher, date

Pagination (main body of pagination only)

Series

Subject headings

Added entries

Local call number

Language (as a fixed field, according to the
MARC II specifications for tag 041)

The level 3 record would be further simplified by omitting all indicators, delimiters, and subfield codes.

This type of record might be useful to libraries that plan to convert their own holdings. The advantage of establishing a minimum standard is that it might promote compatibility among libraries that desire to exchange limited bibliographic records on the same terms.

No matter what level of bibliographic records is produced for the primary conversion operation, truncated records (level 3) might also be available for distribution as an option for potential subscribers. The feasibility of providing this service would depend on the future capabilities of a centralized operation.

In attempting to arrive at any of the three levels described above, it is pertinent to explore the possible effects of a promising technical approach for conversion, which involves no manual pre-editing or only partial editing (cues to the machine) with processing in either case by an automatic format recognition program to assign content designators.[1]

At present, it is not possible to say how successfully this program will perform. Records produced by this method might conceivably be equivalent to level 2 if the full character string were input. On the other hand, the most efficient combination of man and machine effort may not permit assignment of all of the indicators and subfield codes in level 2.

Format recognition is now being studied by the Library of Congress in connection with the MARC Distribution Service for current cataloging data. The effort is being concentrated on use of the machine to assist in the editing process, i.e., partial editing with format recognition to arrive at a level 1 record.

An analysis of the functions of content designators specified in the MARC II format has been made by the Library of Congress in relation to the following functions:

1. Organization of data either for machine segmentation of like categories of information (by date, country, language, etc.) or for human-readable display.
2. Alphabetical filing for the printing of book-form catalogs.

---

[1] Cf. chapter 5, section A4, and appendix G.

3. Searching for an individual item.
4. Retrieval of items by specified arguments.
5. Statistics for management control.
6. Maintenance (updating and control) of data elements in a system.
7. Output of a variety of products (i.e., catalog cards, special listing, machine-readable data, etc.).

The effects of any loss of precision resulting from use of a format recognition program will have to be evaluated in the light of the functions listed above. For example, many indicators and subfield codes are used principally to facilitate programming to produce sophisticated filing arrangements. They add significantly to the complexity of manual editing of MARC II records and may present unsolvable problems for a format recognition program. Whether the benefits of the filing arrangements are worth the cost of achieving them is a legitimate question. For this reason, the Library of Congress and other libraries are re-examining the basic requirements for file arrangement.

If a centralized conversion project does come into being, the cost of conversion to a MARC II record might influence the decision in favor of a record with simpler content designators. The consequences of reducing costs by this means must be weighed against possible disadvantages of a mixed data base at a central source. The supposed savings may be largely offset if future library operations necessitate wholesale revision to achieve a uniform level of machine coding in the entire data base.

During the course of this study, it became more and more evident that a mixed data base (i.e., conversion at different levels) at a central source would be a serious mistake. To avoid this difficulty, it seems desirable to strive for an optimum format for both current and retrospective records by a judicious balance between human and machine assignment of content designators.

Appendix G

FORMAT RECOGNITION

A. Editing as a Factor in Format Recognition

In the context of this study the purpose of a format recognition program is to accept magnetic tape records that have been converted into machine-readable form by some input device and automatically to reconstruct and tag the records according to the specifications of a MARC II record (see appendix F). The working task force considered both alternatives for input devices and alternatives in the amount of human editing (tagging, delimiting, etc.) that would be performed upon the record prior to input. These latter alternatives were defined as (1) full editing: editor assigns all tags, delimiters, etc., prior to conversion to machine-readable form, (2) partial editing: editor assigns a subset of tags, delimiters, etc., prior to conversion, and (3) no editing by a human prior to conversion.

Full editing does not require any format recognition program since the function has been performed completely prior to conversion. Partial editing and no editing both require format recognition of varying degrees of complexity assuming the final product in both instances is a MARC II record. Before an accurate measure of the ideal balance between man and machine can be known, it will be necessary to make a statistical analysis of the characteristics of cataloging records in a variety of languages and an evaluation of the logic of the software that is not only required but possible.

An unedited magnetic tape record can be the result of direct-read OCR or keying by an input device. In the case of the use of a keying device, function codes will be input by the typist to simulate type faces and indentions in the original data and provide the same level of cues as would result from reading the LC printed card by the direct-read OCR. It is obvious that if a keying device were used, some simplified editing could be accomplished at transcription time. For discussion purposes, however, this fine distinction leads to too many variables. Therefore, the format recognition problem for both types of devices is assumed to be the same.

The discussion that describes the conversion of the LC record set by use of direct-read OCR and followed by format recognition is confined to the LC card printed since 1949. Before 1949, the card had three different printing formats. Although the three earlier formats were not substantially different from the cards printed since 1949, the format recognition program would require modification of this interpretation.

Partially edited magnetic tape records would result from some level of editing by a human being followed by transcription by a keying device. Partial editing should result in a more accurate performance by the format recognition program. Given some cues, the machine would make fewer mistakes than if the program were assigned the entire responsibility for the editing function. If a large number of records have to be recycled through the machine because of format recognition errors, processing without pre-editing is highly questionable. It is expensive not only in terms of machine time but, even more important, in terms of the manpower required to proof, correct, and re-key.

If a data element were not identified in a partially edited record, the format recognition logic will be confronted with the same situation as in an unedited record. Since partial editing cannot be defined at this time, it is difficult to make a clear distinction between the two categories when giving examples.

B. Format Recognition Logic

This section gives a brief and over-simplified description of format recognition logic and the problems inherent in this attempt to minimize the human editorial function. The discussion is based on work performed by the Library of Congress with contractual support. Although much thought has been given to format recognition, the work to date is not at a point where it is safe to derive absolute conclusions about its efficiency.

Program algorithms for both partially edited and unedited records would depend on patterns of punctuation, spacing, capitalization, position (right margin, left margin), and type face. In other words, the physical attributes of the printing yield cues to many of the content designators: for example, on LC cards bold type usually signifies the main entry, indention marks the beginning of the fields such as the title, and the LC card number is in the lower right-hand corner of the card.

There are, of course, significant limitations to the capabilities of this technique. First, it is virtually impossible to identify a data element whose sole cue lies in the meaning of the character string itself. For example, it would be difficult to identify the type of subject, i.e., geographic, topical, or political jurisdiction. Given the term Andes, there is no way for the machine to determine the type of subject heading. It would not be feasible to have a lookup table the length of the Columbia Lippincott Gazetteer to identify geographic names. A similar problem exists in distinguishing general subject subdivisions from geographic subdivisions.

Second, the visually discernible printing cues are not always present even for those content designators that can be related to the cues and sometimes, even when present, they are ambiguous. For example, the edition statement is not always separated from the imprint statement by the use of a period; in some cases, a closed bracket is substituted for the period.

1. Main Entry

A name used as a main entry might be identified by format recognition logic without cues by using the following algorithm.

The first recognition problem in analyzing the name main entry would be in determining if, in fact, there was a name entry or if the work was entered under title. This might be determined by the design of an algorithm depending on spacing. (Direct-read OCR under program control can record the spacing on a printed card as characters coded as blanks or spaces.) When a main entry is a name, it begins at the far left margin, about 3/4 inches from the edge of the card. If the name runs over one line, the next line would be printed 15/16 inches from the edge. The title begins a new line indented approximately 1-1/6 inches from the edge. The rest of the body of the entry is printed 15/16 inches from the edge. When a record is entered under title, the card is usually printed in the hanging indention format. The title begins at the far left margin, about 3/4 inches from the edge and each line in the body of the entry following would be printed 15/16 inches from the edge. Under ALA rules, a record entered under title was sometimes printed in paragraph format. In this case, the title main entry could be recognized from the fact that the first line begins 1-1/16 inches from the edge.

Therefore, if the recognition program were dependent on position (spacing) it would be necessary for the computer to "look ahead" at the rest of the title paragraph to distinguish between a name main entry and a title entry.

It might be possible also to distinguish elements in the main entry by type face. The following patterns of 10-point bold, italic, and roman type are used on LC cards.

Title main entry

Elements	Type
Title with an initial article	Roman/bold/roman
Title without an initial article	Bold/roman

Name main entry

Elements	Type
Personal name, title, date, relator	Bold/italic/roman/italic
Personal name, title, date	Bold/italic/roman
Personal name, title or relator; Corporate name, qualifiers or subdivisions	Bold/italic
Personal name, date, relator	Bold/roman/italic
Personal name, date	Bold/roman
Personal or corporate name	Bold

An algorithm could be formulated to scan the characters in the record (equivalent to the first line on the printed card) searching for roman type. If the characters in the roman string were numeric, an assumption could be made that the numerics equaled the date of a name main entry. If the roman type encountered in the first line were alphabetic, a title entry could be assumed.

If the entry were a name entry, the format recognition logic would have to categorize the name into one of many types such as personal name, single surname; personal name, forename; corporate name entered under place, etc.

The program logic for this analysis would be complex. For illustrative purposes, a possible subroutine for the recognition and delimiting of a single personal surname is described below:

a. If the first word is followed by a comma, the name is assumed to be a personal name, single surname. (The possible error in this logic is that the entry might be a place name followed by a comma, e.g., Washington, D. C.)

b. The character string is then searched for a second comma and the data after the second comma is divided into subfields using the following algorithms.

   (1) If the data is numeric, the subfield is assumed to be date, and field is delimited with the date subfield code.

   (2) If the data is alphabetic, the characters are compared against a lookup table of the most common terms used as relators, e.g., ed., comp., illus., etc. If a match occurs, the subfield is delimited with the relator subfield code.

   (3) If no match occurs in point b above, the subfield is considered to be a title subfield and so delimited.

c. The process continues searching for a third comma and a fourth comma if present, recycling through the same subroutine described in b (1)-(3) above. For names not analyzed as personal names beginning with a single surname, other algorithms would be designed to match against keywords or symbols. For example, the words "conference," "symposium," "congress," etc., would usually indicate that the name was that of a meeting. If a period was found following the first word, the name would probably be a corporate name entered under place. A hyphen embedded in the first word usually indicates a personal name beginning with a multiple surname.

It should be noted that it is highly improbable that all types of entry could be recognized by format logic. Those that could not be identified could be tagged as unknown or perhaps erroneously tagged and would have to be corrected in the proofing process.

If some degree of "partial editing" were assumed, the format recognition would be simpler to construct and more accurate in performance. For example, if each major field were to be identified, the logic could concern itself with the indicators and the subfield codes required for the field. In the main entry field, it would be fairly simple to have an editor distinguish between name and title main entries. In addition, the name main entries might be distinguished as personal name, corporate name, meeting, and uniform title. This determination is for the most part simple but occasionally can be troublesome, as in the case of foreign geographic names and corporate bodies.

If the type of main entry is known, the analysis now breaks down into a determination of the kind of name (such as personal name single surname) and, within the name, the pertinent subfields. For personal name single surname, it would be possible to use the logic that depends on the location of the comma after the first word. Since the determination would already have been made that the field contained a personal name, the problem of differentiating between personal name forenames and corporate names would be eliminated. Also eliminated would be the confusion between personal name single surnames and corporate names entered under place when the place was followed by the state or country.

2. Call Number Field

The call number field lends itself readily to automatic format recognition without prior editing. The field could be identified by its position in the lower left-hand corner of the card. (This equates to some position based on spacing in the record.) The presence or absence of square brackets surrounding the call number would determine if the book were in the LC collection. The separation of the call number into class number and book number would be somewhat more difficult, but (based on a sample of 531 call numbers) the following algorithms could insert the delimiter correctly about 94 percent of the time. The delimiter would be placed before the last uppercase alphabetic character unless the last uppercase alphabetic character was preceded by a period. Then the delimiter would be inserted before the period. (Examples: HE355.A3‡A5155 and QC4331‡.L65).

3. Title Field

The title field would be very difficult to divide into its component parts by machine without human assistance. Simple identification of the end of the field would be difficult since the title transcription is frequently made up of several segments separated by periods. Without some partial editing, it would be difficult to separate the end of the title statement from the edition statement. Within the title statement, the problem exists of separating the data into short title, remainder of title, and remainder of title page transcription subfields.

In a small sample of 258 titles, trial algorithms were used with the following results. When a delimiter was inserted after the first mark of punctuation, the short title was distinguished correctly only 77 percent of the time. Attempts to distinguish the remainder of the title page transcription were based on location of the cue word "by." The characters immediately preceding "by" were searched for a comma, a closed bracket, and one of the following words: edited, compiled, translated, preface, introduction, illustrated, prepared, selected, or foreword, and a delimiter was inserted before the word. This algorithm was correct only 76 percent of the time. This rough sample indicates that for maximum efficiency it might be necessary to pre-edit the title field.

4. Author/Title Fields

Another field that would be difficult to analyze by machine is the author/title entry used as a subject entry or as a general added entry. An algorithm that would effectively separate the subordinate units of a corporate name from a following title would probably be impossible to construct and some kind of partial editing would be mandatory.

C. Conclusion

It is not within the scope of this appendix to give a field-by-field analysis of the LC catalog record from the standpoint of format recognition. The studies currently in progress at the Library of Congress indicate that partial editing combined with format recognition processing is a promising alternative to full editing. Figure G.1 is an unedited record on a MARC worksheet. Figure G.2 illustrates the same record partially edited along the lines of the ongoing investigations. Figure G.3 shows this record after full editing. These figures serve to illustrate the degree of human involvement in full editing as opposed to partial editing.

Figure G.1--Unedited MARC worksheet

LIBRARY OF CONGRESS

Information Systems Office
MARC II INPUT WORKSHEET

Edited by ______________________

Date ______________________

Languages LAN ☐			
FFD	Govt Pub 1.	Conf/Meeting 2.	
Festschrift 3.		Index 4.	M E in body 5.
Publisher is M E 6.		Juvenile 10.	Fiction 11.
Biography 12.		Subject is M E 13.	
Pub Date Key 20.		Date 1 21.	
Date 2 22.		Country of Pub 23.	
Repro Form 25.		Contents Forms 26.	
Bib Level 27.		Modified Record 28.	

11-19 (rev 9/68)

Seminar on the Organization and Handling of Bibliographic Records by Computer, *Newcastle-upon-Tyne, 1967.*

Organization and handling of bibliographic records by computer: ⌊proceedings of a seminar sponsored by the Computing Laboratory and the Library of the University of Newcastle-upon-Tyne⌋; edited by Nigel S. M. Cox and Michael W. Grose. Newcastle-upon-Tyne, Oriel P., 1967.

xv, 192 p. facsims., tables, diagrs. 31 cm. 65/-

(SBN 85362 000 8) (B 67-19914)

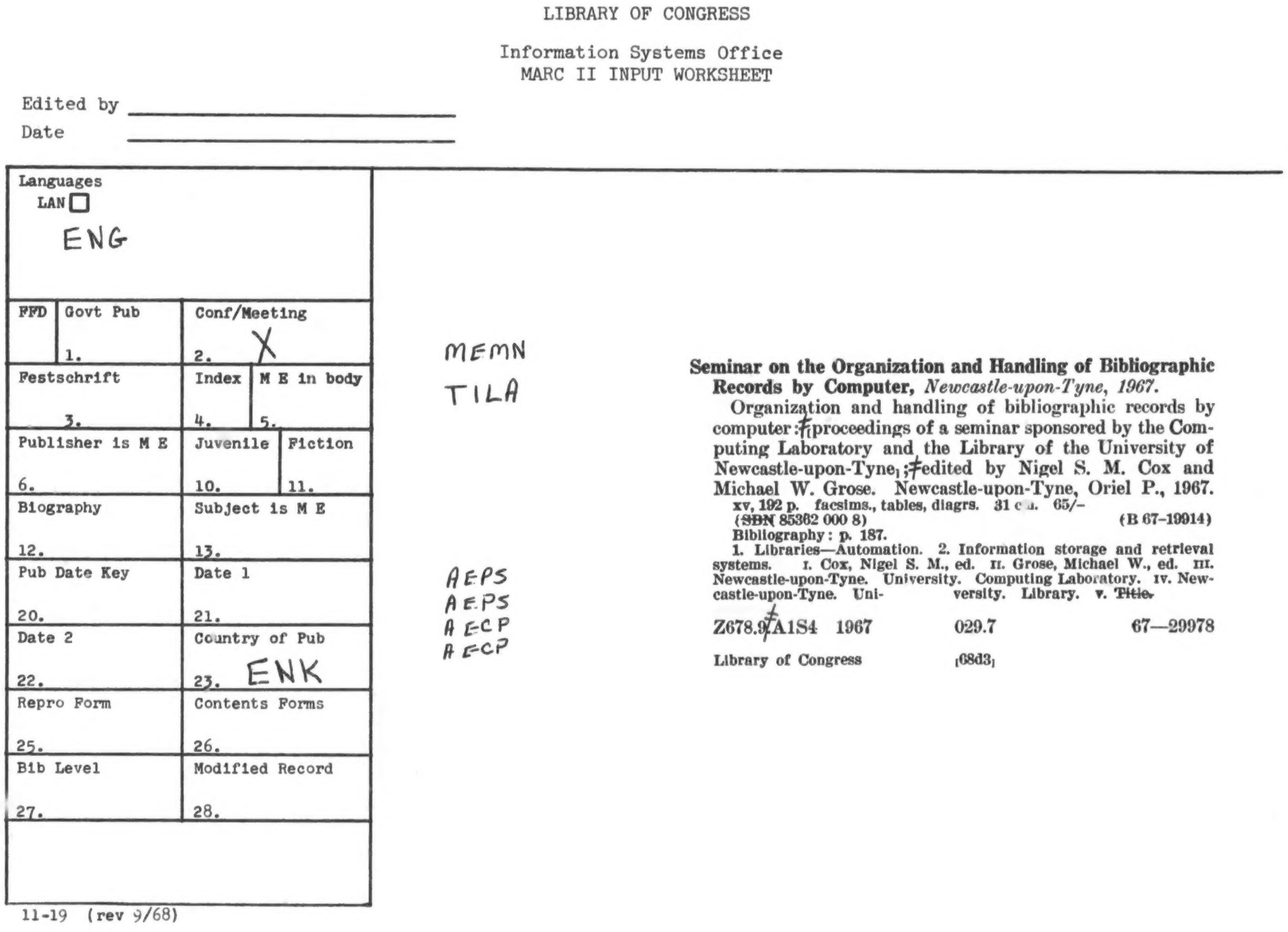
Bibliography: p. 187.

1. Libraries—Automation. 2. Information storage and retrieval systems. I. Cox, Nigel S. M., ed. II. Grose, Michael W., ed. III. Newcastle-upon-Tyne. University. Computing Laboratory. IV. Newcastle-upon-Tyne. University. Library. V. Title.

Z678.9.A1S4 1967 029.7 67—29978

Library of Congress ⌊68d3⌋

Figure G.2--Partially edited MARC worksheet



LIBRARY OF CONGRESS

Information Systems Office
MARC II INPUT WORKSHEET

Edited by ____________________
Date ____________________

Languages LAN☐ ENG		
FFD Govt Pub 1.	Conf/Meeting 2. X	
Festschrift 3.	Index 4.	M E in body 5.
Publisher is M E 6.	Juvenile 10.	Fiction 11.
Biography 12.	Subject is M E 13.	
Pub Date Key 20.	Date 1 21.	
Date 2 22.	Country of Pub 23. ENK	
Repro Form 25.	Contents Forms 26.	
Bib Level 27.	Modified Record 28.	

11-19 (rev 9/68)

MEMN
TILA

AEPS
AEPS
AECP
AECP

Seminar on the Organization and Handling of Bibliographic Records by Computer, *Newcastle-upon-Tyne, 1967.*

Organization and handling of bibliographic records by computer :‡[proceedings of a seminar sponsored by the Computing Laboratory and the Library of the University of Newcastle-upon-Tyne] ;‡edited by Nigel S. M. Cox and Michael W. Grose. Newcastle-upon-Tyne, Oriel P., 1967.

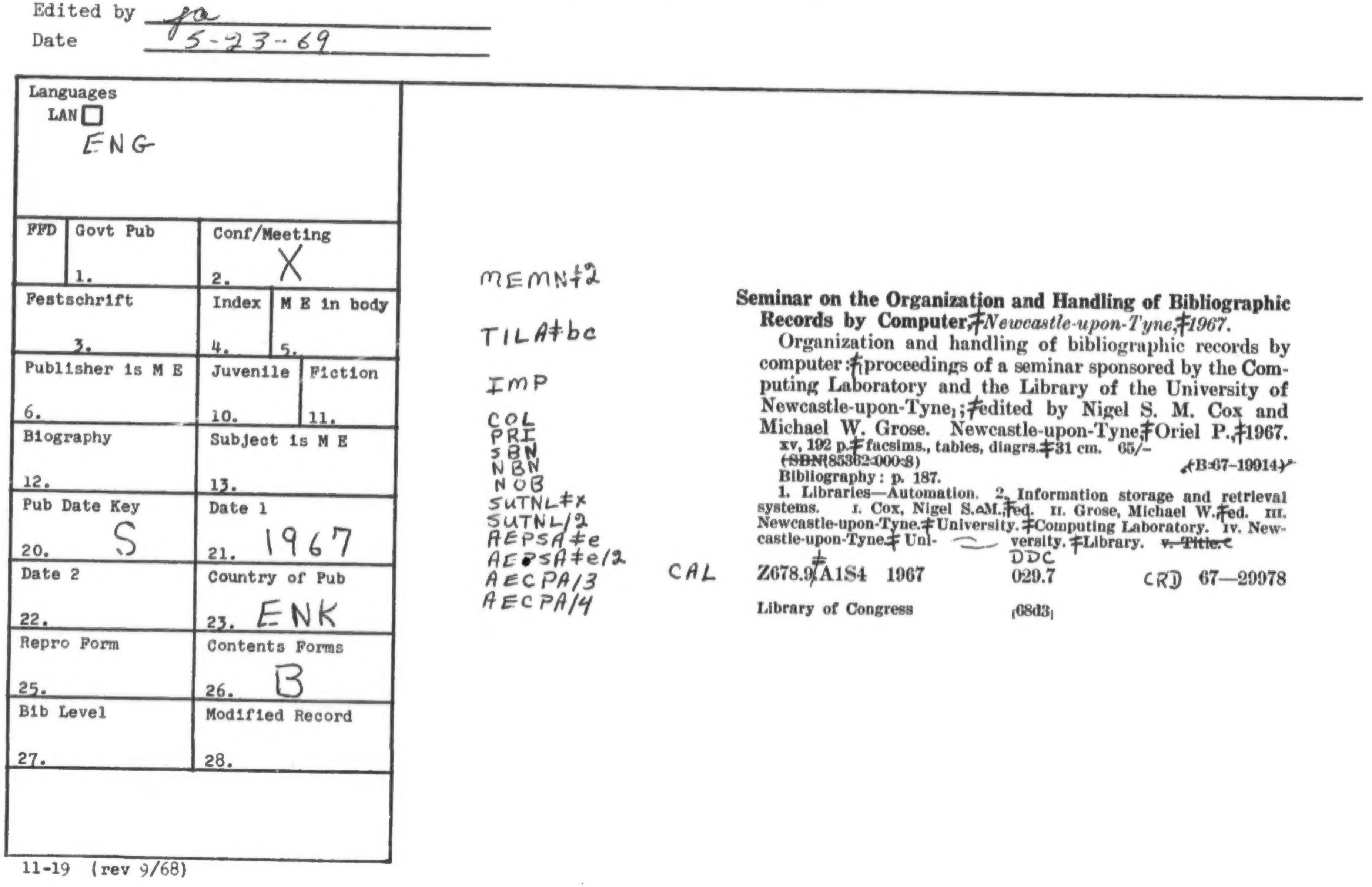
xv, 192 p. facsims., tables, diagrs. 31 cm. 65/-
(~~SBN~~ 85362 000 8) (B 67-19914)
Bibliography: p. 187.

1. Libraries—Automation. 2. Information storage and retrieval systems. I. Cox, Nigel S. M., ed. II. Grose, Michael W., ed. III. Newcastle-upon-Tyne. University. Computing Laboratory. IV. Newcastle-upon-Tyne. University. Library. V. ~~Title.~~

Z678.9‡A1S4 1967 029.7 67—29978

Library of Congress [68d3]

Figure G.3--Fully edited MARC worksheet



LIBRARY OF CONGRESS

Information Systems Office
MARC II INPUT WORKSHEET

Edited by ja
Date 5-23-69

Languages LAN ☐ ENG			
FFD	Govt Pub 1.	Conf/Meeting 2. X	
Festschrift 3.		Index 4.	M E in body 5.
Publisher is M E 6.		Juvenile 10.	Fiction 11.
Biography 12.		Subject is M E 13.	
Pub Date Key 20. S		Date 1 21. 1967	
Date 2 22.		Country of Pub 23. ENK	
Repro Form 25.		Contents Forms 26. B	
Bib Level 27.		Modified Record 28.	

11-19 (rev 9/68)

MEMN‡2
TILA‡bc
IMP
COL
PRI
SBN
NBN
NOB
SUTNL‡x
SUTNL/2
AEPSA‡e
AEPSA‡e/2
AECPA/3
AECPA/4

CAL

Seminar on the Organization and Handling of Bibliographic Records by Computer,‡*Newcastle-upon-Tyne,‡1967.*
Organization and handling of bibliographic records by computer:‡proceedings of a seminar sponsored by the Computing Laboratory and the Library of the University of Newcastle-upon-Tyne;‡edited by Nigel S. M. Cox and Michael W. Grose. Newcastle-upon-Tyne‡Oriel P.,‡1967.
xv, 192 p.‡facsims., tables, diagrs.‡31 cm. 65/-
~~(SBN~~85362-000-8) ~~(B67-19914)~~
Bibliography: p. 187.
1. Libraries—Automation. 2. Information storage and retrieval systems. I. Cox, Nigel S.△M.,‡ed. II. Grose, Michael W.‡ed. III. Newcastle-upon-Tyne.‡University.‡Computing Laboratory. IV. Newcastle-upon-Tyne‡University.‡Library. ~~V. Title.~~
DDC
Z678.9‡A1S4 1967 029.7 CRD 67—29978
Library of Congress [68d3]

Appendix H

COMPUTER REQUIREMENTS FOR A NATIONAL BIBLIOGRAPHIC SERVICE

A. Introduction

This appendix presents an analysis of the hardware and software requirements to provide machine-readable bibliographic information to the library community from a central source. The service would be designed to provide magnetic tapes containing blocks of records in selected categories on a subscription basis and to satisfy on-demand requests for specific records.

The postulated time frame for this effort is as follows: design of the system by 1970; site preparation and implementation of system by 1972; and conversion of records and initiation of the distribution service in the period 1972-1976. Thereafter, conversion of current cataloging and any other retrospective records that might be appropriate would supply material for a continuing service. Additional hardware would be required if and when the data base exceeded the size allowed by the capacities of the present design.

Volumes, production rates, and cost figures have been obtained by extrapolation from current data. Much of this information stems from the MARC Distribution Service which has many similarities to the projected service. Assumptions and estimates have been kept as realistic as possible; if anything, they are pessimistic. This preliminary system design was constructed for the present report as a model for estimating cost, time, and performance. A definitive design would require one or two man-years of detailed analysis.

B. Distribution Services

1. General

The function of the central installation would be to convert bibliographic records to machine-readable form, to maintain them in a central store, and to make them available to the library community. Designing a centralized system for distributing machine-readable records for retrospective material poses many problems. The regular production of records over a period of years would make a subscription service possible. At a regular interval (perhaps weekly) a magnetic tape containing newly converted records could be distributed to subscribers. Since few potential users will require all of the records if they cover a wide range of languages and dates, some means should be found to satisfy their varying needs. The following patterns of service might be considered:

a. Complete sets of tapes to libraries, regional processing centers, and commercial services that desire to search against a complete file.

b. Subsets of the total file by major language category (e.g., English; other roman alphabet languages) and/or date (possibly limited to 10-year periods).

c. On-demand service by Library of Congress card number or author/title.

2. On-Demand Service

The on-demand capability would allow customers to order specific records already in machine-readable form either by LC card number or by author and title. On-demand requests would result in the accumulation of records extracted for a customer from the total data base using either or both accesses. The records selected for a customer would be distributed on magnetic tape.

The on-demand capability is conceptually feasible but its achievement requires a great deal of planning and design. A small number of on-demand requests (2,000 per day) has been used in this report to provide the basis for estimates for this type of service. Note that the term on-demand request is not envisaged to mean on-line requests for the time period 1972-1976.

C. Hardware Requirements

1. Computer Configuration

The central installation should include a medium-scale, third-generation computer with 8-bit byte handling capabilities. It would probably not be critical to have on-line capabilities because the installation would operate in a batch-processing mode. Since many of the processes are input/output bound, however, the operating system should have multiprogramming capabilities for efficient use of the main frame.

The computer should include the standard peripheral devices: card reader, card punch, and line printer. The printer should have at least a 600-line-per-minute rate and be able to print 132 print positions per line.

There should be at least six magnetic tape drives which permit sorting with a two-way merge. The drives should be 60 KC drives (800 bits per inch, 75 inches per second). Additional tape drives would be useful not only to sort more efficiently but also to duplicate tapes for the distribution services. It would be highly desirable to be able to read tapes forward and backward.

Two classes of mass storage would be required. A relatively fast access disk pack device (50-100ms average access time) would be required for a directory to the data base (author/title index). The IBM 2314 disk or its equivalent (roughly 200-million bytes of storage) would be suitable. A large-scale, less rapid access device (100-200ms average access time) would be needed for storage of the records themselves. An example of this kind of device is the Bryant 4000-series disk with 400-million bytes capacity. Total storage capacity can be expanded by additional units.

The rental for a computer with the above characteristics, exclusive of the disk devices, is in the range of $25,000-$35,000 per month. Examples of such computers are the SDS Sigma 7, RCA Spectra 70/45, and the

IBM 360/50. The cost of the disks varies with the number of records to be maintained in the data base. The disk costs were based on the following assumptions:

a. The average length of a record is 500 bytes including overhead characters for machine manipulation.[1] Each large-scale 400-million-byte disk would hold approximately 750,000 records, allowing for some nonusable space (see figure H.1).

b. The faster disk packs would be used for the author/title index and for the entry to the threaded list structure. In the worst case, this would require 13-million bytes of fixed overhead for the index plus 40 bytes per record. Therefore, each disk pack of 29-million bytes would accommodate 40-byte overhead fields for 700,000 records. The exception would be the first pack which could accommodate only 40-byte overhead fields for 400,000 records because 13-million bytes on this disk would have to be used for fixed overhead area for the author/title index. Allowance was made for nonusable space. (See section D4 for details of the access method.)

c. The rental for the IBM 2314 disk is $5,410 per month, not including an additional $20 per disk pack per month, and the rental for the Bryant 400-million-byte disk is $8,350 per month. The later figure includes an estimate of maintenance cost, whereas maintenance on the IBM disk was not included as it is handled separately on a time-and-materials basis.

---

[1] Based on analysis of 391 records on the MARC II test tape. The shortest record had 281 characters, the longest 1,074.

Figure H.2 illustrates the combined costs of the series of two different disk devices needed to maintain the data base and the author/title index. Figure H.3 illustrates how the cost per 1,000 records would vary with the total number of records. The saw-tooth curve represents the sum of the monthly rentals of the Bryant disk and the IBM 2314; the addition of each Bryant disk represents a large step function, while an additional 2314 adds a small step function. Neither figure H.2 nor figure H.3 allow for the temporary utilization of surplus space on the 2314 for data base records until another Bryant disk is required.

D. Software Requirements

1. General

The general software requirements of the system would be those of any data processing computer installation: operating system, assembler, compilers, dumps, utilities, sort/merge, etc. Most of this software should be supplied by the vendor. In addition, special service programs would be needed for customer accounting, subscription list maintenance, mailing list generation, etc.

The programs designed especially for the creation, maintenance, and retrieval aspects of the system would all be of considerable complexity. They fall into three general processing subsystems:

Figure H.1--Storage capacity of large-scale disks, in terms of number of records stored

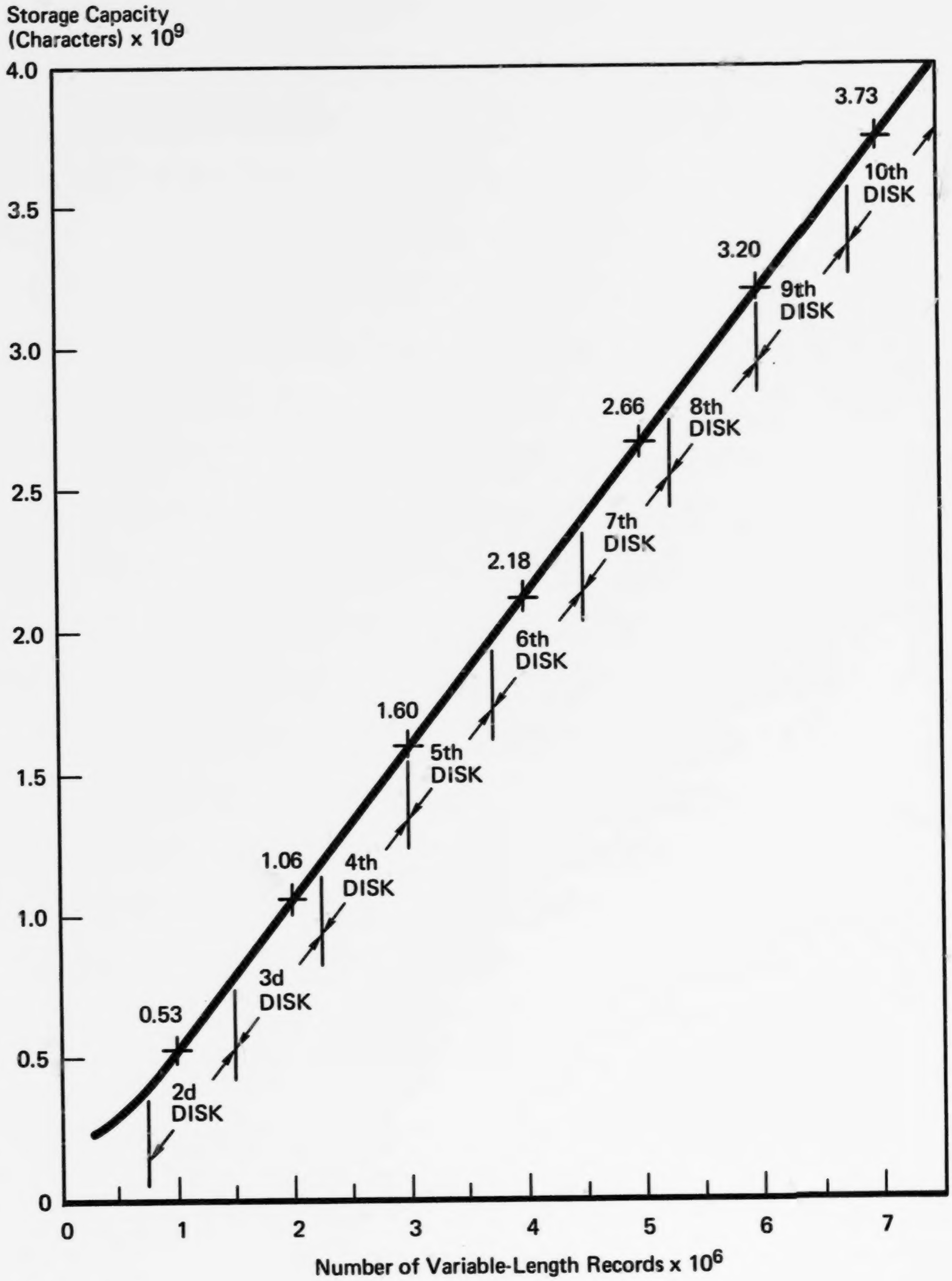
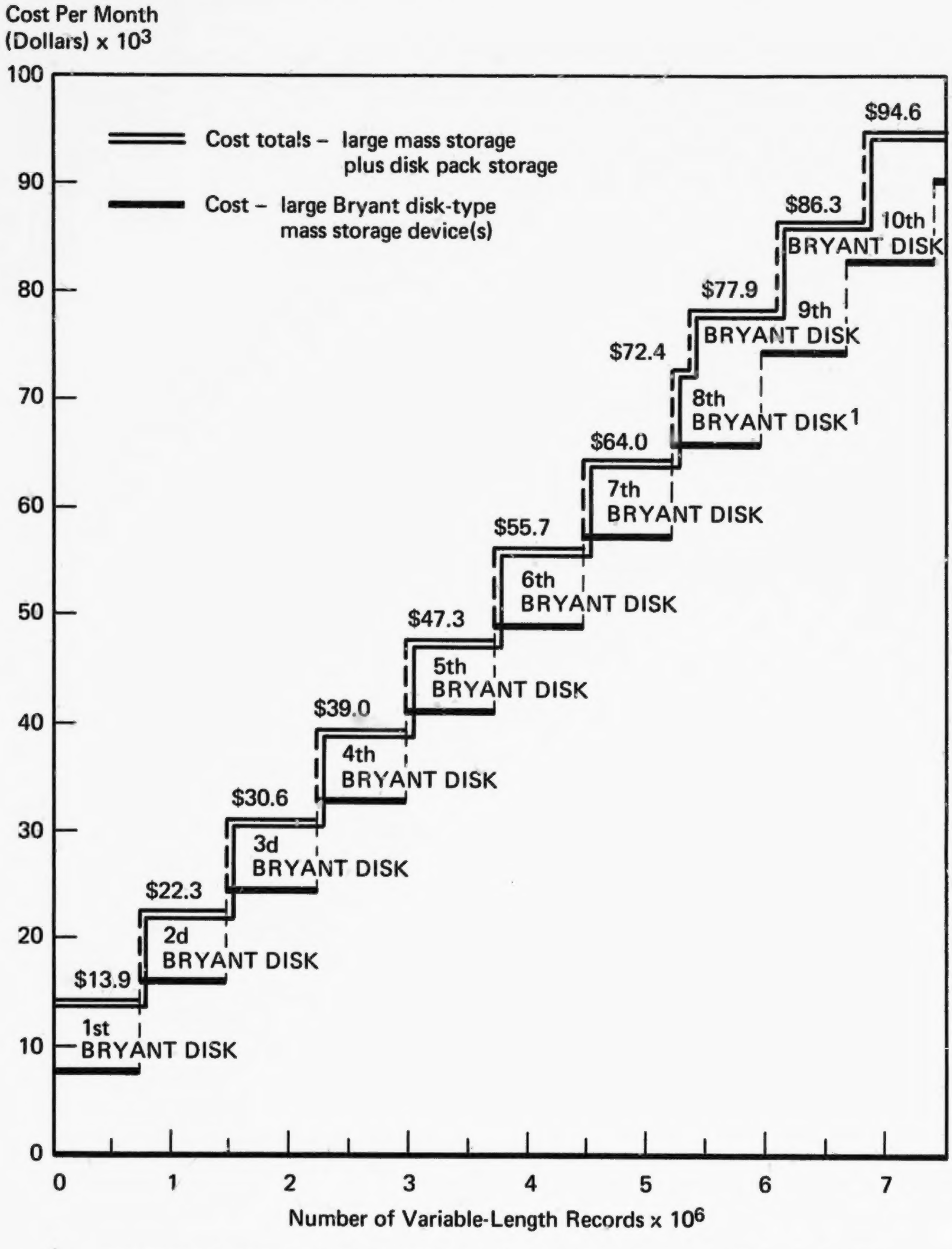


Figure H.2--Monthly cost of devices to store data base and index in terms of number of records stored



[1] THE COST OF THE FIRST 2314-TYPE DISK IS INCLUDED IN THE COST OF FIRST SEVEN BRYANT DISKS. WHEN MORE THAN SEVEN BRYANT DISKS ARE USED, A SECOND 2314 WILL BE REQUIRED; HENCE THE SHARP INCREASE IN COST.

Figure H.3--Cost of storage per 1,000 records, by number of records stored

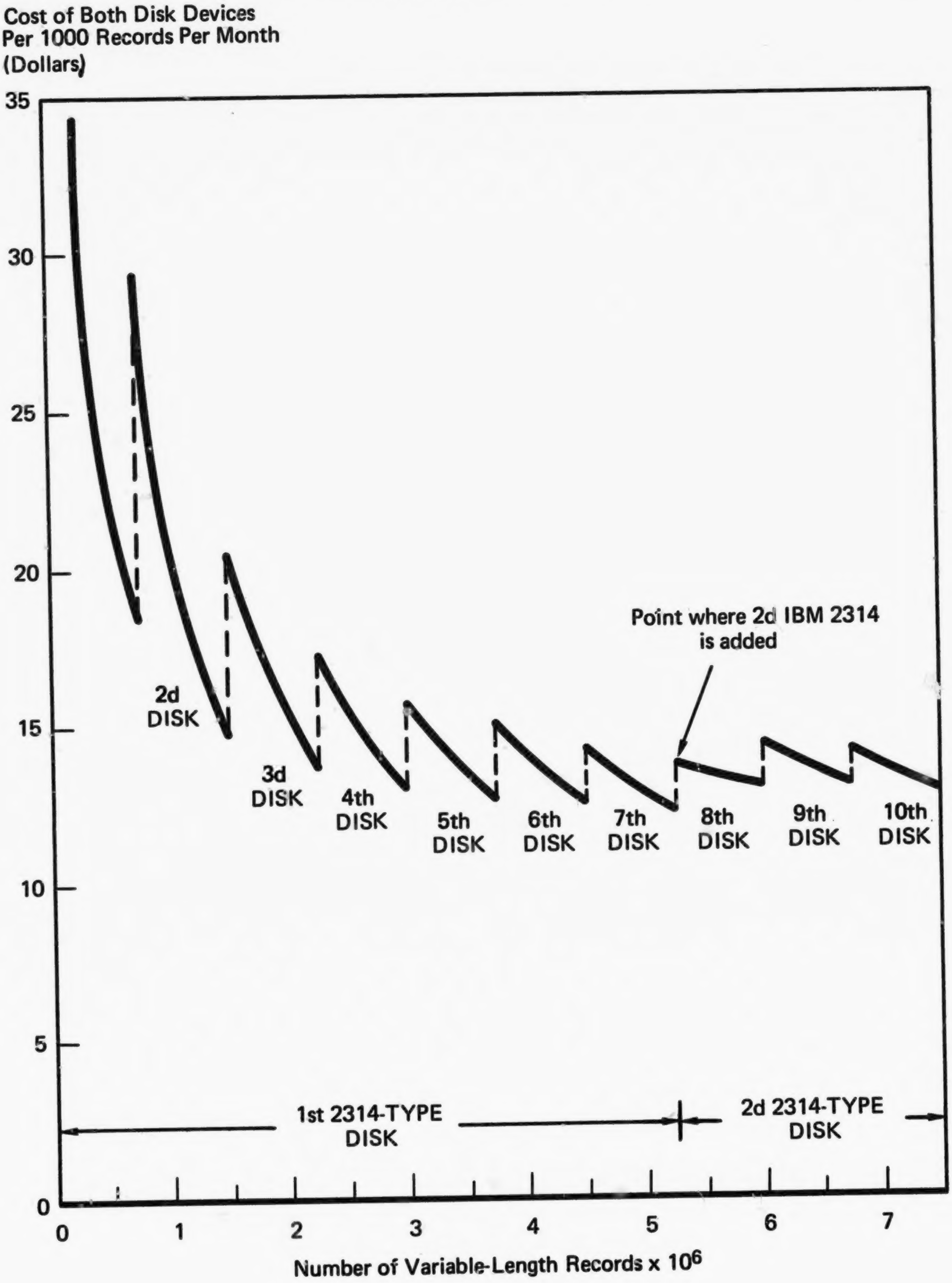
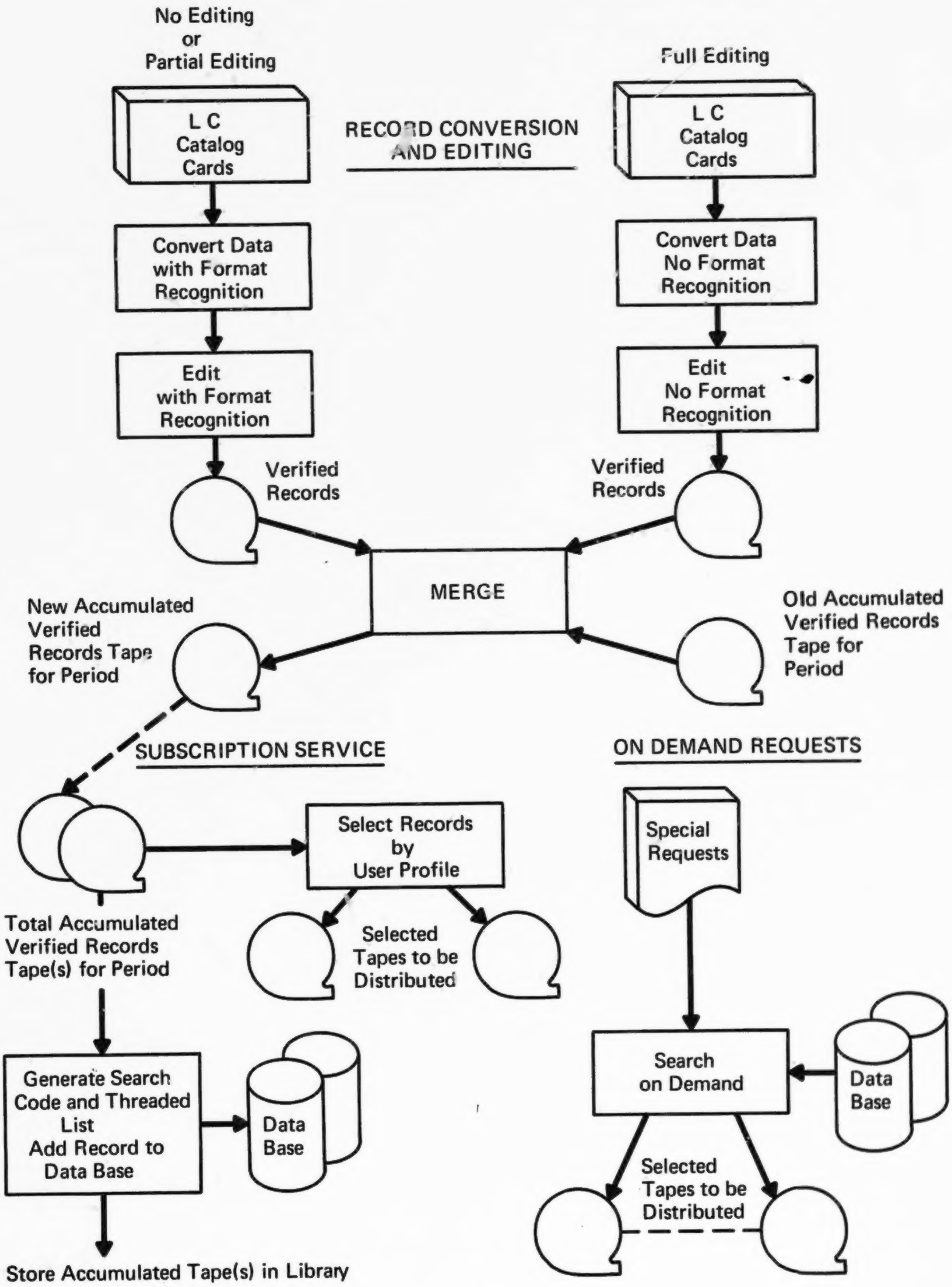


Figure H.4--System for a projected national bibliographic service


No Editing or Partial Editing
L C Catalog Cards
Convert Data with Format Recognition
Edit with Format Recognition
Verified Records
RECORD CONVERSION AND EDITING
Full Editing
L C Catalog Cards
Convert Data No Format Recognition
Edit No Format Recognition
Verified Records
MERGE
New Accumulated Verified Records Tape for Period
Old Accumulated Verified Records Tape for Period
SUBSCRIPTION SERVICE
Select Records by User Profile
Total Accumulated Verified Records Tape(s) for Period
Selected Tapes to be Distributed
Generate Search Code and Threaded List Add Record to Data Base
Data Base
Store Accumulated Tape(s) in Library
ON DEMAND REQUESTS
Special Requests
Search on Demand
Data Base
Selected Tapes to be Distributed

Record Conversion and Editing Subsystem

Perform format recognition (OCR or keyboard transcription; no editing or partial editing).

Edit and format (keyboard transcription; full editing).

Check validity.

Produce formatted print.

Perform file maintenance (new, corrected, or verified records).

Subscription Service Subsystem

Select records by user profile.

Duplicate selected records.

Data-Base-Related Subsystem

Generate search code and threaded lists; add record to data base.

Search on-demand.

The interrelation of these programs is shown in figure H.4.

2. Record Conversion and Editing Subsystem

a. Perform Format Recognition (OCR or keyboard transcription; no editing or partial editing)

The format recognition module would accept magnetic tape records that had received no editing or had been partially edited prior to input and would automatically analyze the data to convert the record into a tagged formatted internal processing record (see appendix G for a description of format recognition).

It is apparent that a program of considerable complexity would be required to analyze records to the same degree of definition as is now attained entirely by human editing (see figure H.5).

b. Edit and Format (keyboard transcription; full editing)

The edit and format module would accept records that have been fully edited prior to input and transform the input format to the internal processing format. All tags, indicators, delimiters, etc., would be specified by an editor and input at conversion time (see figure H.6).

c. Check Validity

This program would check the records for content consistency and correctness, and would flag all machine-detectable errors to call them to the attention of the editors during proofing. This program would be used for both modules specified in a and b above (see figures H.5 and H.6).

d. Produce Formatted Print

This program would accept bibliographic records and produce formatted printouts for proofing and correction. The program would be used for the modules specified in a and b above.

Consideration must be given to the printing of records in a data base containing a variety of alphabets. Since even English language records may contain words in nonroman alphabets, the Library of Congress had to face this problem for the MARC Distribution Service (for English language monographic cataloging data.) It was decided that the nonroman alphabets would be romanized until time permits a detailed analysis of the required character sets and the associated problems of input, manipulation, and

Figure H.5--Subsystem for record conversion and editing with format recognition

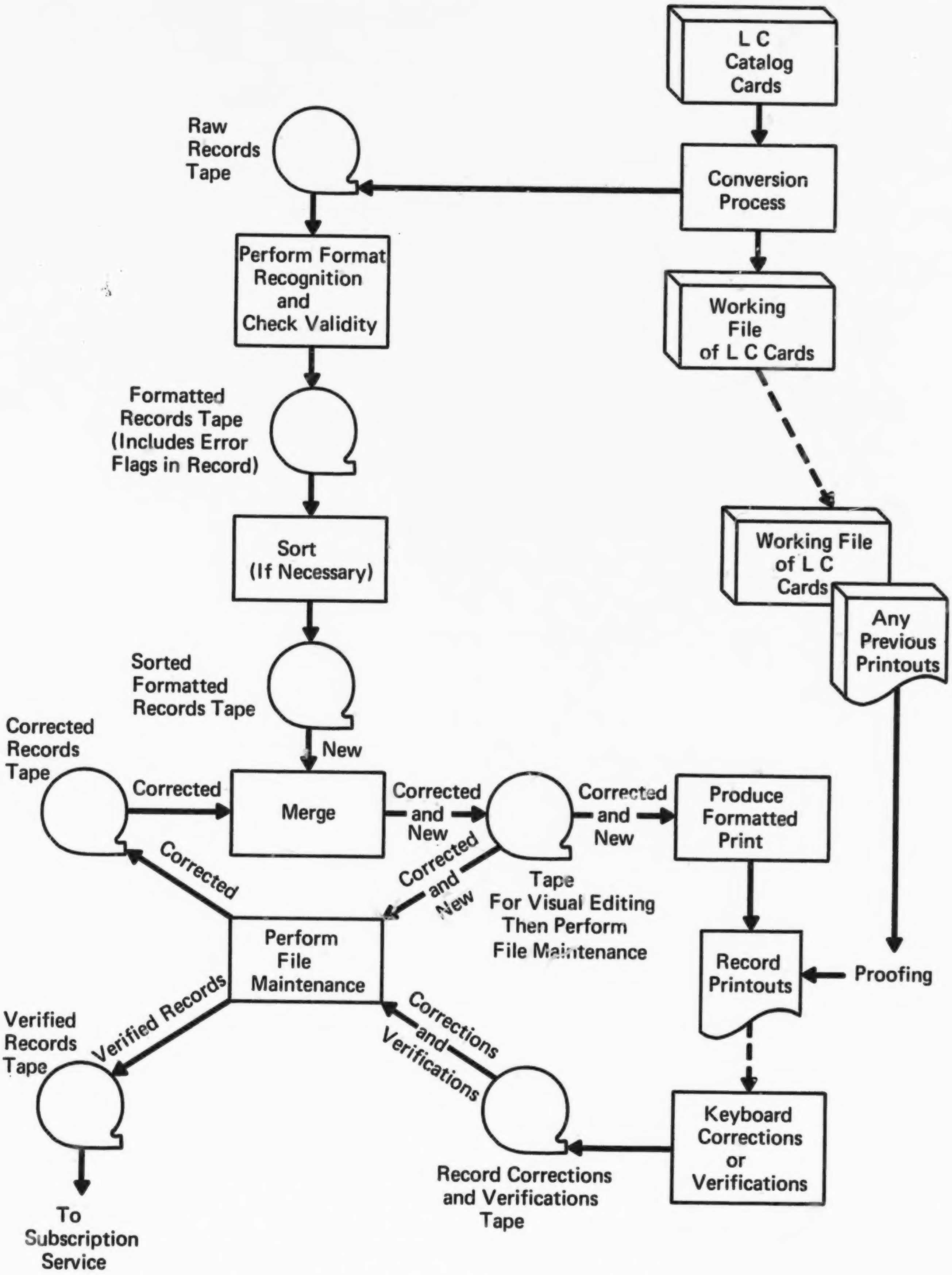
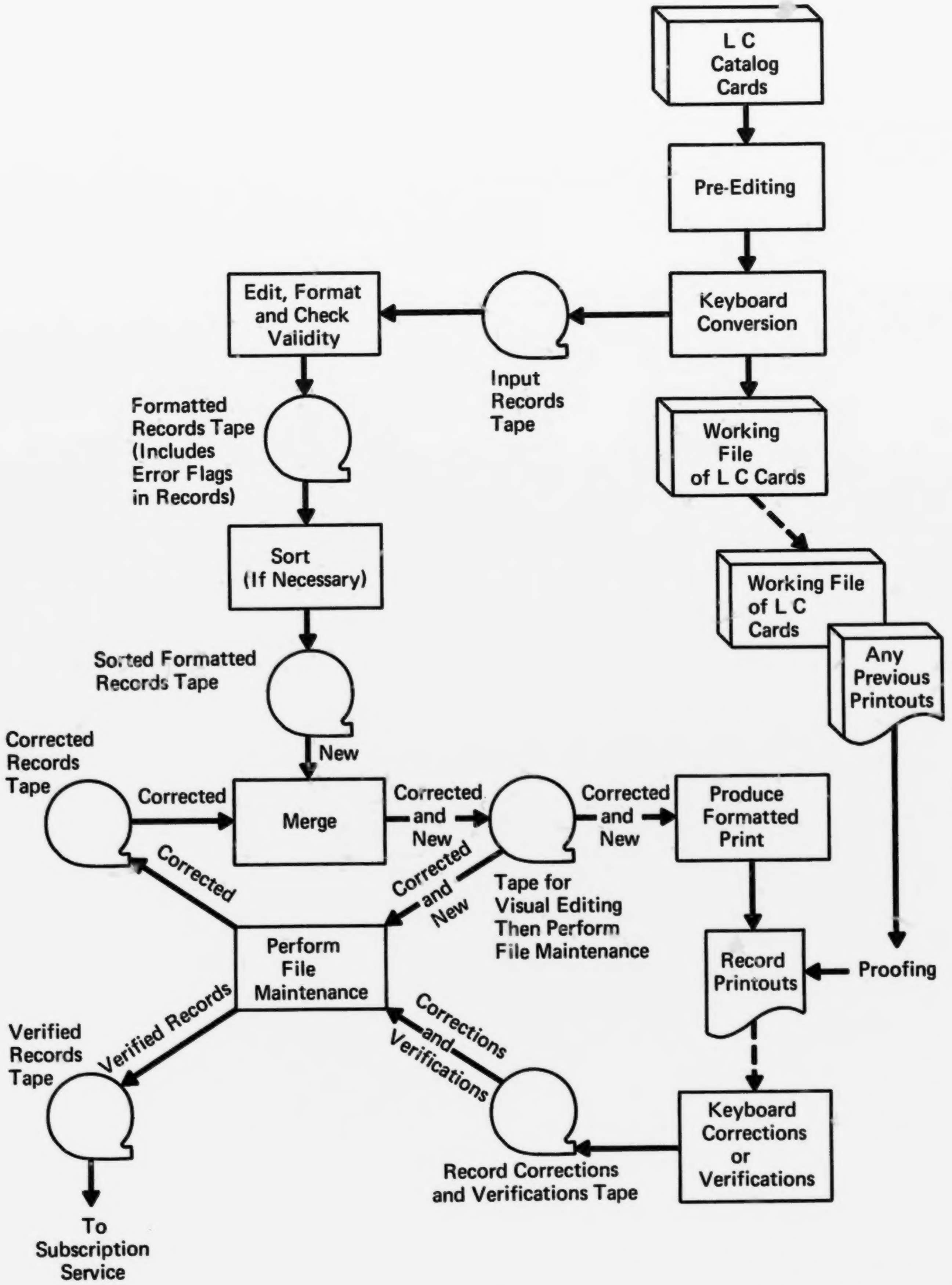

L C Catalog Cards
Conversion Process
Raw Records Tape
Perform Format Recognition and Check Validity
Working File of L C Cards
Formatted Records Tape (Includes Error Flags in Record)
Sort (If Necessary)
Working File of L C Cards
Any Previous Printouts
Sorted Formatted Records Tape
New
Corrected Records Tape
Corrected
Merge
Corrected and New
Corrected and New
Produce Formatted Print
Corrected
Corrected and New
Tape For Visual Editing Then Perform File Maintenance
Perform File Maintenance
Record Printouts
Proofing
Verified Records
Verified Records Tape
Corrections and Verifications
Keyboard Corrections or Verifications
Record Corrections and Verifications Tape
To Subscription Service

Figure H.6--Subsystem for record conversion and editing without format recognition


L C Catalog Cards
Pre-Editing
Keyboard Conversion
Input Records Tape
Edit, Format and Check Validity
Working File of L C Cards
Formatted Records Tape (Includes Error Flags in Records)
Working File of L C Cards
Any Previous Printouts
Sort (If Necessary)
Sorted Formatted Records Tape
New
Corrected Records Tape
Corrected
Merge
Corrected and New
Corrected and New
Produce Formatted Print
Corrected
Corrected and New
Tape for Visual Editing Then Perform File Maintenance
Perform File Maintenance
Record Printouts
Proofing
Verified Records
Corrections and Verifications
Verified Records Tape
Keyboard Corrections or Verifications
Record Corrections and Verifications Tape
To Subscription Service

display. Any retrospective conversion project will be faced with similar decisions regarding the nonroman alphabets.

Depending on the data base selected, the possible decision to preserve the vernacular form of a nonroman alphabet, and the desirability of being able to proof character for character, (i.e., the original representation of a character would be preserved in the printed output) different methods of print capability may be postulated. For example, a computer installation could be assumed to have a chain (or train) designed to include the Cyrillic alphabet as well as the roman alphabet. Naturally, since the number of characters of both alphabets would exceed the number of characters of a single alphabet, print speed would be reduced.

If the data base contained more than one nonroman alphabet, a technique to segregate records by alphabet would have to be designed to allow operator intervention to change the chain (or train). On the other hand, an installation might find it expedient to have a chain (or train) limited to the roman alphabet, numerals, and punctuation. The greater number of alphabetic segments would enable the chain to print faster. In this case, if the record contained a diacritic and the character could not be printed, the proofer would have no way of reading and correcting the missing character. In the final analysis, a judgment would have to be made on the basis of cost (in terms of man hours vs. machine hours) as to the most efficient solution to the problem for any given data base (see figures H.5 and H.6).

e. Perform File Maintenance (new, corrected, or verified records)

The file maintenance module would accept new, corrected, or verified bibliographic records. New records would be written on a working tape and a printout would be made for proofing purposes. Corrections would cause the equivalent bibliographic records to be modified and written on a corrected records tape and to be merged with new input in the next editing cycle. The verified records would be written on a verified records tape, which would be merged with the accumulated verified records for this distribution period. This program would be used for both modules specified in a and b above (see figures H.5 and H.6).

3. Subscription Service Subsystem

a. Select Records by User Profile

This program would accept an accumulated verified records tape and generate output tapes of records selected according to user profiles. In addition to the verified records tape, a user profile tape would be used as input. This would have the users' names, addresses, and accounting information, grouped by profile (i.e., the category of record desired). One output of this program would be an updated user profile tape, containing amended accounting information, plus data for any new users, whose profiles could be entered through the card reader.

Assuming six magnetic tapes on the computer, three could contain user profiles so that one pass would suffice for three different profile selections. There would be only two types of profiles: those with only one customer (a unique profile) and those with more than one. For the former case, a mailing label would be printed (or typed) while the tape was generated. For the latter case, the label information could be written on the selected records tape for use by the duplicate selected records program (see figure H.7).

b. Duplicate Selected Records

This program would accept the tape containing selected records from the previous program and generate duplicate copies of them for the appropriate number of users. If six tape drives were used, up to five duplicate tapes might be generated concurrently. The user information in the second file of the input tape would be used to print (or type) mailing labels as the duplicate tapes are written (see figure H.7).

4. Data-Base-Related Subsystem

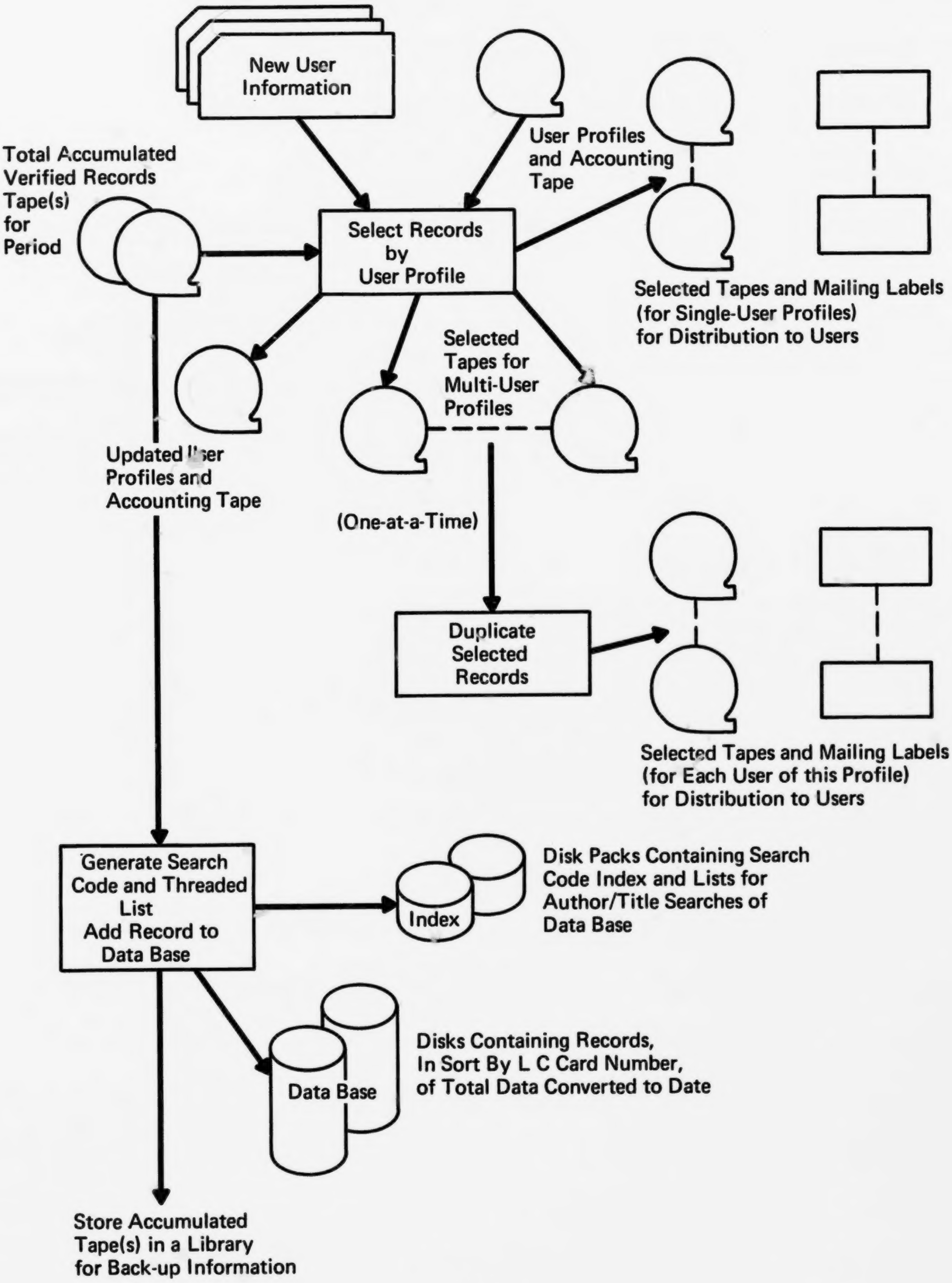
a. Generate Search Code and Threaded List; Add Record to Data Base

The search code referred to in this report involves automatic compression of specified machine-readable data by the method described by Ruecking.2/ The code is constructed by compressing up to four words in a title and up to four more words representing last names of authors for a minimum of two and a maximum of eight four-letter codes. Ruecking claims

---

2. Ruecking, Frederick H., Jr. Bibliographic retrieval from bibliographic input; the hypothesis and construction of a test. Journal of library automation, v. 1, December 1968, 227-238.

Figure H.7--Subsystem for subscription service and generation of search code and threaded list for data base


New User Information
User Profiles and Accounting Tape
Total Accumulated Verified Records Tape(s) for Period
Select Records by User Profile
Selected Tapes and Mailing Labels (for Single-User Profiles) for Distribution to Users
Selected Tapes for Multi-User Profiles
Updated User Profiles and Accounting Tape
(One-at-a-Time)
Duplicate Selected Records
Selected Tapes and Mailing Labels (for Each User of this Profile) for Distribution to Users
Generate Search Code and Threaded List Add Record to Data Base
Index
Disk Packs Containing Search Code Index and Lists for Author/Title Searches of Data Base
Data Base
Disks Containing Records, In Sort By L C Card Number, of Total Data Converted to Date
Store Accumulated Tape(s) in a Library for Back-up Information

a high degree of uniqueness (98-99 percent) in the code resulting from a title. Such a technique might be used to generate an author/title index automatically and to relate it to the LC card number.

Extensive research and testing is required to determine the most efficient system for bibliographic searching. Since this was impossible within the time frame of this study, it was assumed that the search code would be used in a threaded list structure.

The maximum number of four-letter code groups that can result from this scheme can be easily calculated, since the first character may be any letter, the second and third may be any letter or blank, and the fourth may be any consonant or blank. The result is 26 x 27 x 27 x 21= 398,034[3]. Even though some of the combinations are unlikely, the scheme assumes all of them are possible and an index is generated on a disk pack consisting of 32 bytes for each code group of the 398,034. This will require over 12 million bytes of storage. A rounded figure of 13 million bytes has been used for all calculations in the present report. This figure is less than one half of the capacity of one 2314 disk pack. The advantage of generating all possibilities would be that the index (hereafter referred to as a permanent index), once created, would be fixed; that is, it would never need to be shifted because new records were added. Even more important, the position of the 32-byte field for a given code group could be directly calculated from the code group itself and searching would not be necessary.

---

[3] This assumes that diacritical marks and special characters in roman alphabet languages are disregarded.

The 32 bytes for a code group in the permanent index would consist of eight 4-byte links pointing to threaded lists (hereafter referred to as list entries) containing the LC card numbers of all records with search codes (two to eight groups) that contained the group in that code position on the disk.

A threaded list is a classic form of file organization used to access records from keys. In its simplest form, there are two groups of data: a key directory and records. Typically, the key directory, contains an attribute (name, code, or abbreviation), the address of the first record in storage possessing that attribute, and usually the list length (i.e., the total number of records that are referenced in the full list). The record will usually contain a major data subset and a series of links. Each link is associated with a particular key and is a pointer to a subsequent record also associated with the same key. There can be as many links associated with a record as there are keys associated with that record. The pointing from key directory to record, from record link to subsequent record, and from subsequent record link onward is called threading, and there will be as many threads as links as keys. For example, link 1 of a possible eight links for a record for which the title compresses to AMER would link to the LC card numbers of all records, for which the title compressed to the code group AMER.

Given the permanent index, only a list entry (i.e., an entry to the threaded list structure) would be needed for each record added to the data base. This list entry would consist of the LC card number, a flag byte, and two to eight 4-byte links to connect the entry in the list structure. It is assumed that seven bytes would be enough to contain a card number; the year and serial number can be expressed in packed decimal in four bytes, the alphabetic prefix, expressed in three bytes.[4] The flag byte would signify which links were present. Thus, if an author/title generated two codes for the title and one for the author, this byte would have the pattern 11001000_2.

There would only be one 4-byte link for every search code group generated from the author/title(s) of the record. Therefore, 40 bytes per record for this entry (7 plus 1 plus 32 [8 x 4]) would be the worst-case condition for overhead. In fact, this seems an extremely unlikely occurrence, since it would only occur for a title having four or more significant words in its title and four or more authors. However, this worst-case figure of 40 bytes overhead per record was used in volume projections.

Given the above, the program to build the search code for a new record would extract the LC card number and construct the search code from the author/title (this could be done so easily that it might be desirable to carry the search code permanently in the data base record). The code groups would then be used to locate links and the new card number would be linked into the structure.

---

[4] This pattern will also accommodate the new 8-digit LC card number which has no alphabetic prefix.

The other function of this program would be to add the record to the data base (hereafter referred to as main data base) on the larger mass storage device. This could be done in such a way that the records would be in ascending sequence on the LC card numbers. A possible method of referencing the records more efficiently than by a serial search (which would be implied if the records are in ascending order) would be to store the records in partitioned areas of storage according to the range of the number. This technique is sometimes called the "bucket" process. Each partitioned area would be referred to by a range of the numbers involved. This ordering would allow the retrieval of records, using the card number, to be effected using a simple binary search technique.

It should be noted that the permanent index would point to a list entry containing the card number of a record, not the record itself. This would be necessary because, when a record was added to the data base there might not be room to store it in its proper place in card number order (one of the assumptions above). Therefore, the record would be stored where room was available and a reference made to its locations. As the number of these references increased over a period of time, the performance of referencing the data base would be degraded, and so the file should be reorganized periodically to restructure the data base in a more efficient manner. This could be done with impunity as long as the permanent index does not directly reference record positions (see figure H.7).

b. Search On-Demand

This program would be essentially the converse of the previous one. It would allow a selected record to be retrieved by author/title or LC card number.

Given an author/title request, this program could retrieve the card number. This would be accomplished by converting the author/title to a search code and looking up the list entry for each search code group in the permanent index. The links would be traced through the lists to locate a list entry with enough common links to satisfy a threshhold test. The linking could be done so that a simple test would reveal the point where the search had failed thus making it unnecessary to search to the end of every list. The result would be the list entry containing the card number of the record.

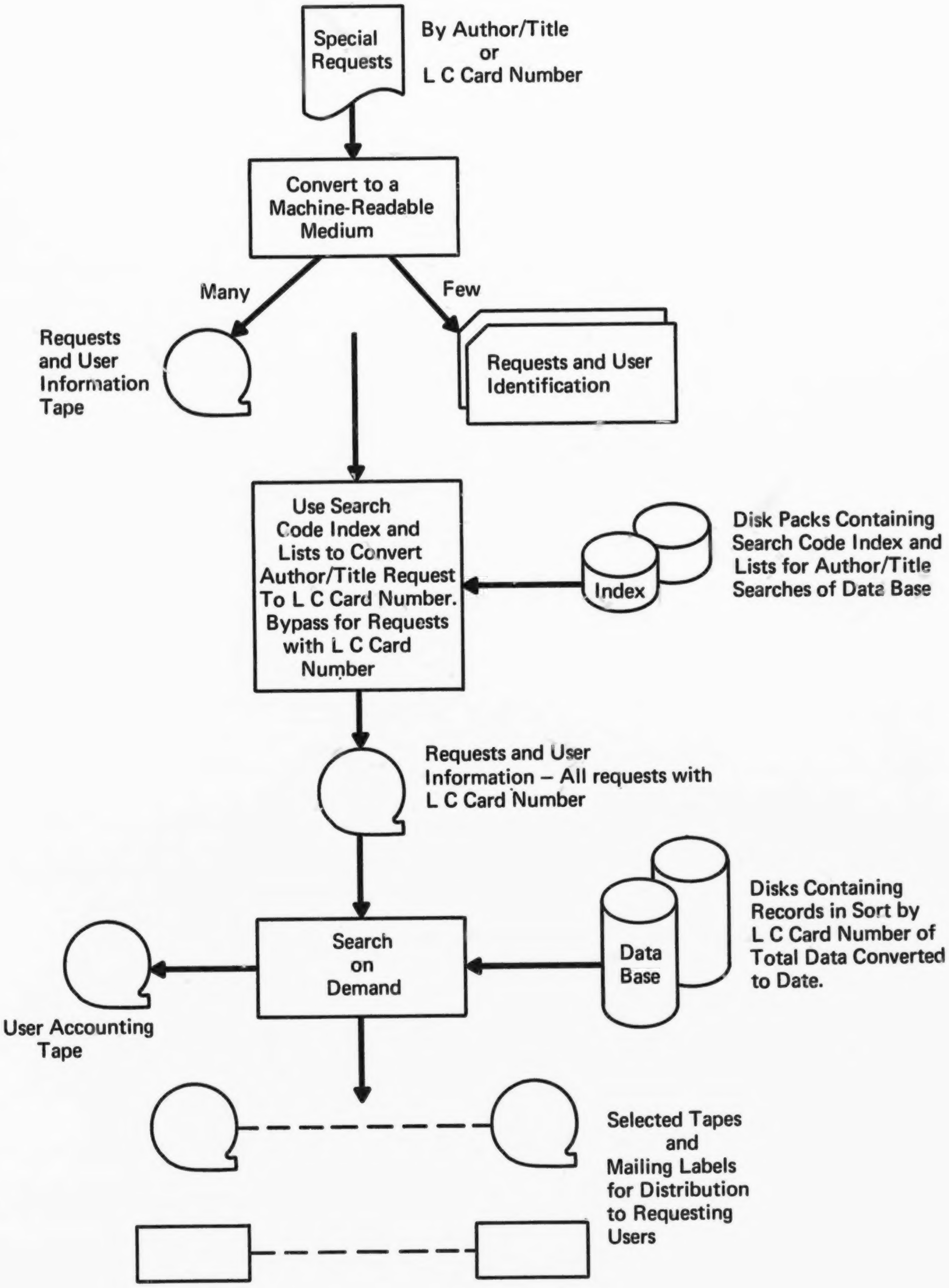
The card number would then be used to retrieve the record exactly as if it were input in the first place. A binary search of "dividing the dictionary" technique could be used. The desired card number would be compared against the number of records in the physical center of the main data base which would have to be in ascending order by LC card number. If the desired number were less than the number at the center of the data base, the next test would be made in the middle of the first half of the data base. If the desired number were greater, the next test would be made in the middle of the bottom half of the data base. This process would continue, halving each time, until the desired number was found.

This technique has the advantage of limiting the number of such tests that must be made. Where 2^n produces a number greater than or equal to the number of records, the maximum number of searches is equal to n. For example, with seven million records, a maximum of only 23 tests would need to be made because $2^{22} = 4,194,302$ and $2^{23} = 8,388,608$. At an average access time of 100 milliseconds on a disk (such as the Bryant 4000-series disk), this would equal a worst-case search time of 2.3 seconds. For convenience, three seconds has been used for timing studies.

The search could be reduced even further by using a table of "milestone" LC card numbers. These would be the card numbers of records occurring at regular intervals in the disk(s); for example, the number of the first record in every sector, every cylinder, etc. Such a table could be built after collecting the numbers by a pass through the disk(s) when the program was initialized. If this were done, a two-level binary search could be constructed, first in the "milestone" table and subsequently, when the disk area of search has been narrowed, in disk storage itself. The advantage to this technique is that a search in a table in memory is virtually instantaneous as compared to a 100-millisecond average disk access. The first few searches would be the most extensive and time consuming if all were made against the disk, thereby biasing considerably the average access time. In the "milestone" table, assuming only 256 (2^8) values, the maximum number of disk searches would be reduced to $23 - 8 = 15$, cutting the total search time to 1.5 seconds (see figure H.8).

The following discussion describes the application of the threaded

Figure H.8--Subsystem for on-demand searches


Special Requests
By Author/Title or L C Card Number
Convert to a Machine-Readable Medium
Many
Few
Requests and User Information Tape
Requests and User Identification
Use Search Code Index and Lists to Convert Author/Title Request To L C Card Number. Bypass for Requests with L C Card Number
Index
Disk Packs Containing Search Code Index and Lists for Author/Title Searches of Data Base
Requests and User Information – All requests with L C Card Number
Search on Demand
Data Base
Disks Containing Records in Sort by L C Card Number of Total Data Converted to Date.
User Accounting Tape
Selected Tapes and Mailing Labels for Distribution to Requesting Users

list structure to the storage and retrieval of bibliographic records for a national bibliographic service.

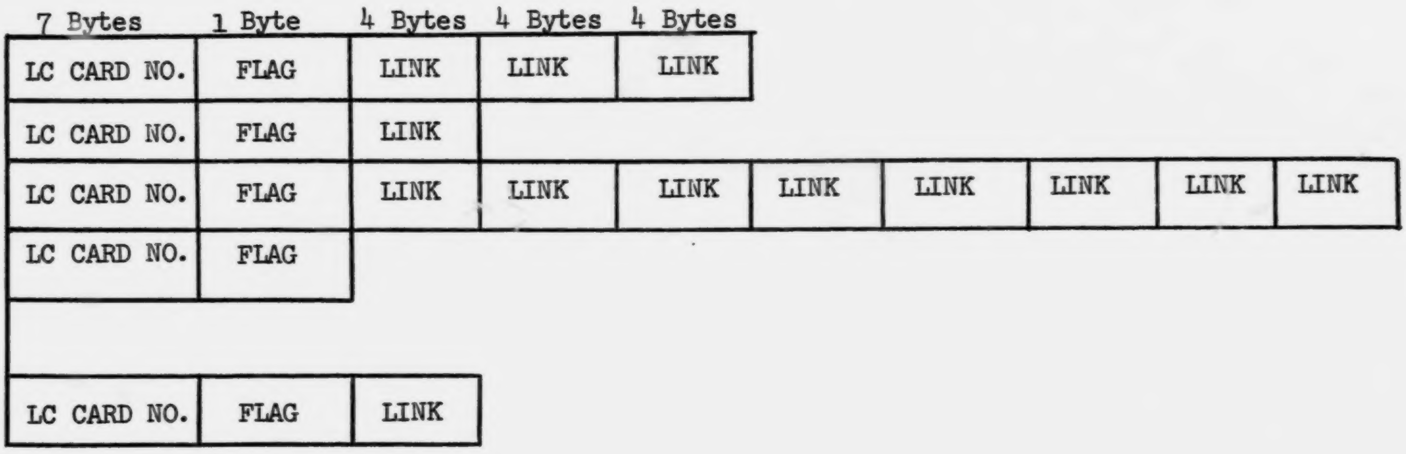
Figure H.9--Diagram of a permanent index

Position

	1	2	3	4	5	6	7	8
Code Group 1 (A)	LINK	LINK	LINK	LINK	LINK	LINK	LINK	LINK
Code Group 2 (AA)	LINK	LINK	LINK	LINK	LINK	LINK	LINK	LINK
Code Group 3 (AAA)	LINK	LINK	LINK	LINK	LINK	LINK	LINK	LINK
Code Group n (AMER)	LINK	LINK	LINK	LINK	LINK	LINK	LINK	LINK
Code Group 398,034 (ZZZZ)	LINK	LINK	LINK	LINK	LINK	LINK	LINK	LINK

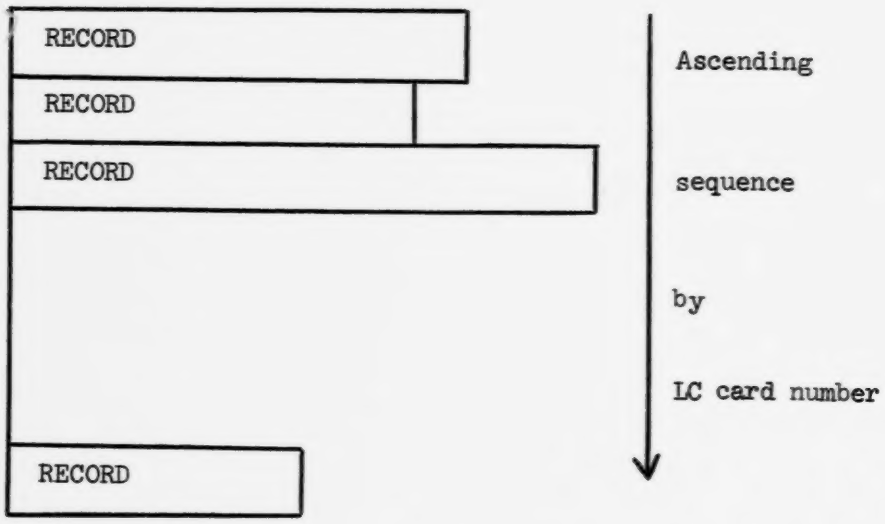
A permanent index consisting of 398,034 sets of eight lists each would be generated. Each link would be a 4-byte pointer which, if non-zero, would contain the address of the first list entry in the threaded list for the specific code group (e.g., AMER) in a specific position in the search code (e.g., position 2). Starting with a title and its author(s), a search code would be constructed containing up to eight alphabetic code groups. In the particular search code assumed, exactly 398,034 different alphabetic code groups are possible. Some possible code groups are A, B, AMER, ZZZZ. Since each alphabetic code group may exist in up to eight positions of the search code, the permanent index permits up to eight links for each code group. For example, link 2 in the set of eight links corresponding to code group AMER would point to the first list entry corresponding to a search code in which AMER exists in position 2. As the permanent index would comprise all possible code groups, the address of the eight-link index entry could be directly computed from the code group without searching.

Figure H.10--Diagram of list entries



Each nonzero link of the permanent index would point to a list entry representing the first case in the list that satisfied the conditions of a given code group in a given position. First occurrence list entries would be pointed to by links in the permanent index, subsequent list entries would be pointed to by links in other list entries. A list entry would consist of an LC card number, a flag, and zero to eight links. The LC card number would be the primary access to the main data base of full records in large mass storage. The 8-bit flag byte would indicate which of the eight possible links (if any) were present. If no links were present, only one record (as represented by its LC card number) would have a search code that satisfied the particular code group in the particular position indicated. The presence of all eight links in a list entry would indicate that the record had eight code groups in its search code. Since two different code groups could not occupy the same position in the search code, each record would be represented by only one list entry, and there would be no link ambiguity. The list entries would have a variable length of eight to 40 bytes.

Figure H.11--Diagram of main data base



The main data base would contain full bibliographic records in ascending sequence by LC card number. The output information from the list entry would be the card number which would be used to locate the full records in the main mass storage. Several methods of locating the record from the card number would be possible: (1) a "binary search" which would eliminate successive halves of storage; (2) a direct search based on a starting location of a specified range of records (the "bucket" approach); or (3) the use of an intermediate directory of record addresses ordered by LC card number.

5. Programming Effort Estimates

The following estimates indicate the magnitude of the programming effort required to design, implement, and checkout the programs described in this appendix.

Program	Man-years
Format Recognition (OCR or keyboard transcription; no editing or partial editing)	3.0
Edit and Format (Keyboard transcription; full editing)	2.0
Formatted Print	.5
Check Validity	.5
Perform File Maintenance (new, corrected, and verified records)	2.0
Select Records by User Profile	1.0
Duplicate Selected Records	.25
Generate Search Code and Threaded List; Add Record to Data Base	2.0
Search On-Demand	1.0
Service Programs	2.0
Total	14.25

On a contractual basis at an estimated $35,000 per man-year, the total programming effort would amount to about $499,000. An in-house effort calculated at $15,000 per man-year would cost approximately $214,000. An in-house effort to complete these programs would probably require a greater elapsed time because of the difficulties in recruiting and retaining qualified programmers.

E. Computer Processing Time

1. Assumptions

The following assumptions have been made in computing the data in this section. They are based largely on present MARC II experience at the Library of Congress on the IBM 360/40 with DOS. Needless to say, operations on a more powerful machine (an IBM 360/50 or comparable equipment) or in a multiprogramming environment would result in different time estimates.

a. The conversion rates for input are assumed to be one to seven million records over a four-year period. Using 208 weeks in four years, this rate is 5,000 to 35,000 records per week or 1,000 to 7,000 per day.

b. Magnetic tape recorded at 800 bits per inch is assumed to hold 20,000, 500-byte records. The time to read or write a full tape at 60 KC is assumed to be six minutes.

c. The number of times a record will cycle through the machine is a function of the type of pre-editing a record received and whether the record was compared with the LC Official Catalog. A full discussion of the factors involved in recycling appears in Section E2.

d. The workloads for the subscription service and on-demand record requests cannot be estimated with a high degree of confidence. On-demand requests have been assumed to be at the rate of 2,000 per day: 30 percent by author/title, the remainder by LC card number.

2. Estimated Processing Rates of Programs

a. Perform Format Recognition for unedited records[5]: 4 seconds per record

b. Perform Format Recognition for partially edited records[5]: 3 seconds per record

c. Edit and Format: 3 seconds per record

The rates for a, b, and c were estimated from the MARC System Pre-Edit/Format Edit/Content Edit programs which require a total of three seconds to process a record. Format recognition for unedited records was considered to be much more complex.

d. Produce Formatted Print: 3.4 seconds per record

e. Perform File Maintenance: 3 seconds per record

f. Generate Search Code and Threaded List; Add Record to Data Base: 6 seconds per record

This estimate was based on prior experience in index building programs.

g. Search On-Demand: 3.9 seconds per record

This was considered to take approximately the same processing time as does Generate Search Code and Threaded List; Add Record to Data Base. Half of the time should be spent searching the search code structure and half in retrieving the record. It was assumed that search on-demand requests would break down according to present LC Card Division experience:

---

[5] Assumes a validity checking process by a common program.

30 percent by author/title and 70 percent by LC card number. The time for an author/title search was assumed to be 6 seconds/record; 3 seconds/record was assumed for an LC card number search. Therefore, an average time of 3.9 seconds was used for estimating search on-demand per record.

h. Sort/Merge (including preprocessing[6]) for printing records to be compared against the LC Official Catalog

Many techniques for internal sorting are available: exchanging, insertion, shell exchange, counting, P-operations, and others. A particular strategy can be chosen as most efficient if (1) special data characteristics have been analyzed, (2) file size is known, and (3) certain hardware techniques are used. Manufacturer sorting software takes one or more of these factors into account, but it does not allow a change in strategy for each program execution.

The amount of available core directly affects the size and the number of strings that will be developed by the internal sort.

The following assumptions have been made to complete sort/merge time:

1. 65,536 addressable bytes of memory.
2. The buffering capability of one selector-channel with IBM S/360 DOS (estimate based on MARC System experience).

---

6. Preprocessing is a pass executed prior to the sort/merge to build a sort key that can be used to approximate library filing order. The calculations for preprocessing time are based on the MARC system experience.

3. Access speeds for third-generation equipment.
4. Undefined records to the preprocessor; variable records input to sort; undefined records output for sort.
5. Little or no inherent sequencing exists in input.
6. One sort key of four to 10 characters in length.

Table H.1--Preprocessing and sort time for specified numbers of records

Number of input records (in thousands)	Time (in minutes)[1]		
	Sort	Pre-processor	Total
2	4.5	1.5	6
5	6	3	9
7	9.5	4.2	13.7
10	13	5	18
20	33	8	41
30	49	14	63
40	65	20	85
50	81	26	107
60	107	36	143
70	125	42	167
80	142	48	190
90	[2]	54	-

1. Set-up time is not included.
2. 81,780 records is maximum for the configuration assumed for the table.

There are a number of interrelated variables affecting this process. Memory size affects the internal sort that is chosen. The sorting technique affects the length of the strings that are produced. The size of available core affects the string length. The string length determines the number of strings. The amount of data affects the number of strings. The number of strings determines the most advantageous merging technique. The best merging technique is dependent on the number of tape units and on the original sort technique used. An additional complicating factor is that the number of records that can be kept in memory varies with record size. The time-estimates were obtained from various formulas modified by experience with processing of MARC II records.

i. Select Records by User Profile and Duplicate Selected Records: 6 minutes per tape

The processing rates for these two programs are considered to be magnetic tape input/output bound. The rate for a full tape (20,000 records): is six minutes. This figure was used consistently to calculate the run times for various numbers of records. Actually, in a real situation, the processing times for larger numbers of records might be somewhat reduced by duplicating more than one tape at a time.

3. Recycling of Records

To calculate machine running times for the technical alternatives described in chapter 6, it was necessary to make certain a priori estimates about the percentage of records that would contain errors because the format recognition program would assign incorrect content designators. These errors would be corrected by the human editor during proofing. The correction would be re-keyed and recycled through the machine system to correct the machine-readable data base. In addition, regardless of the type of pre-editing given the record and the performance of the format recognition program, some editing and keying errors would occur under all conditions both in original editing and keying of the record and reediting and re-keying. Therefore, for calculation purposes, the following assumptions were made:

a. Fifty percent of records receiving full pre-editing will be rejected for incorrect tagging, keying errors, etc., during the first proofing process.

b. One hundred percent of the unedited records processed by format recognition will be rejected during the first proofing process.

c. Sixty percent of the partially edited records processed by format recognition will be rejected during the first proofing process.

d. Ten percent of records edited and re-keyed after proofing will be rejected during each proofing process after the first.

In addition to these assumptions allowance had to be made for the percentage of otherwise acceptable records that would recycle because of changes made when they were compared with the Official Catalog. On the assumption that catalog comparison would result in an average of 20-percent change across the board, the 50-percent reject rate was raised to 60 percent (the 50 percent rejected plus 20 percent of the 50 percent accepted) and the 60-percent reject rate to 68 percent (the 60 percent rejected plus 20 percent of the 40 percent accepted).

No measure was made relative to the number of errors per record; that is, one error in a record is considered a reject record equal to a reject record with many errors.

The number of records in the machine editing cycle at any one time consists of the following:

a. New records

b. Records corrected and re-keyed from the previous day's new records

c. Sum of all records from previous days still in the system which have been recorrected and re-keyed.

The total number of records in the cycle after the first pass can be expressed as a summation of terms in a geometric progression: $a + ar + ar^2 \ldots\ldots = \frac{a}{1-r}$ where a is the number of rejections after the initial cycle, and r (the number of rejections after each subsequent cycle) is less than one.

Let n = number of new records per day.

p = percentage of new records rejected on the first pass of of records through editing cycle and re-entered on the second day.

a = np = number of rejects input for a second pass through the machine editing cycle.

r = .1

Therefore, summation of all records in cycle, $\Sigma = n + \frac{np}{1-.1} = n\left(1+\frac{p}{.9}\right) = n\,(1 + 1.11p)$. The reject rates for all possible conditions are as follows:

(1) No editing, no comparison with Official Catalog: 100 percent reject rate from first proofing.

(2) No editing, comparison with Official Catalog: 100 percent reject rate from first proofing.

(3) Partial editing, no comparison with Official Catalog: 60 percent reject rate from first proofing.

(4) Partial editing, comparison with Official Catalog: 68 percent reject rate from first proofing.

(5) Full editing, no comparison with Official Catalog: 50 percent reject rate from first proofing.

(6) Full editing, comparison with Official Catalog: 60 percent reject rate from first proofing.

(1) and (2) are the same since the assumption of 100 percent reject rate due to no editing cannot be adjusted to a higher percentage to reflect the 20 percent change caused by the comparison with the Official Catalog, i.e., the 20 percent is subsumed by the 100 percent.

Assuming 1,000 records a day:

(1) and (2) for $p = 1.00$, $\Sigma = 1{,}000[1+1.11(1)] = 2{,}110$

(3) for $p = .6$, $\Sigma = 1{,}000[1+1.11(.6)] = 1{,}666$

(4) for $p = .68$, $\Sigma = 1{,}000[1+1.11(.68)] = 1{,}755$

(5) for $p = .5$, $\Sigma = 1{,}000[1+1.11(.5)] = 1{,}555$

(6) for $p = .6, \Sigma = 1{,}000[1+1.11(.6)] = 1{,}666$

4. Processing Times

Tables H.2 and H.3 show the computer processing times for input by various technical alternatives to produce 1,000 to 7,000 new records per day. Table H.4 shows the computer time for performing maintenance and service functions on a weekly basis at different production levels. The limit of the system would be reached at a daily conversion rate of about 5,000 new cataloging records. This would amount to approximately one million records a year and the maximum of five million records would be reached in about five years. At 5,000 records a day, the computer processing time would approach 24 hours per day and a larger computer or a second computer would be required.

Table H.2--Computer processing times (hours and minutes) for conversion of 1,000 new records per day, by type of pre-editing

Function	Processing time in seconds/record	No editing				Partial editing				Full editing			
		Without catalog comparison		With catalog comparison		Without catalog comparison		With catalog comparison		Without catalog comparison		With catalog comparison	
		Number of records	Time	Number of records	Time	Number of records	Time	Number of records	Time	Number of records	Time	Number of records	Time
Format recognition or Edit and format	3 or 4[1]	1,000	1:07	1,000	1:07	1,000	0:50	1,000	0:50	1,000	0:50	1,000	0:50
Sort	.12 or .3[2]	1,000	0:02	1,000	0:05	1,000	0:02	1,000	0:05	1,000	0:02	1,000	0:05
Print[3]	3.4 or 3.8[4]	2,110	2:00	2,110	2:06	1,666	1:34	1,755	1:46	1,555	1:28	1,666	1:41
Perform file maintenance[3]	3	2,110	1:46	2,110	1:46	1,666	1:23	1,755	1:28	1,555	1:18	1,666	1:23
Total			4:55		5:04		3:49		4:09		3:38		3:59

1. Format recognition applied to an unedited record requires 4 seconds. When it is applied to a partially edited record, the task is less complex and requires only 3 seconds. Editing and formatting of a fully edited record requires 3 seconds.
2. The sort by LC card number for records without catalog comparison requires .12 seconds per record. The addition of the alpha-sort for records with catalog comparison increases the time to .3 seconds per record. The rates are not exactly linear.
3. This function involves new and corrected records in the system on any one day. See section E3 for a discussion of recycling.
4. The print time for records without catalog comparison is 3.4 seconds per record. The time for printing records with catalog comparison in the two-up format is 3.8 seconds per record.

Table H.3--Daily computer processing times

Type of editing and processing function	Processing times (hours and minutes) for specified numbers of records converted per day						
	1,000	2,000	3,000	4,000	5,000	6,000	7,000
No editing; without catalog comparison							
Record conversion and editing	4:55	9:50	14:45	19:40	24:35	29:30	34:25
Search on demand[1]	2:10	2:10	2:10	2:10	2:10	2:10	2:10
Total	7:05	12:00	16:55	21:50	26:45	31:40	36:35
No editing; with catalog comparison							
Record conversion and editing	5:04	10:08	15:12	20:16	25:20	30:24	35:28
Search on demand[1]	2:10	2:10	2:10	2:10	2:10	2:10	2:10
Total	7:14	12:18	17:22	22:26	27:30	32:34	37:38
Partial editing; without catalog comparison							
Record conversion and editing	3:49	7:38	11:27	15:16	19:05	22:54	26:43
Search on demand[1]	2:10	2:10	2:10	2:10	2:10	2:10	2:10
Total	5:59	9:48	13:37	17:26	21:15	25:04	28:53
Partial editing; with catalog comparison							
Record conversion and editing	4:09	8:18	12:27	16:36	20:45	24:54	29:03
Search on demand[1]	2:10	2:10	2:10	2:10	2:10	2:10	2:10
Total	6:19	10:28	14:37	18:46	22:55	27:04	31:13
Full editing; without catalog comparison							
Record conversion and editing	3:38	7:16	10:54	14:32	18:10	21:48	25:26
Search on demand[1]	2:10	2:10	2:10	2:10	2:10	2:10	2:10
Total	5:48	9:26	13:04	16:42	20:20	23:58	27:36
Full editing; with catalog comparison							
Record conversion and editing	3:59	7:58	11:57	15:56	19:55	23:54	27:53
Search on demand[1]	2:10	2:10	2:10	2:10	2:10	2:10	2:10
Total	6:09	10:08	14:07	18:06	22:05	26:04	30:03

1. Based on 2,000 records per day.

Table H.4--Weekly computer processing times for specified functions, by number of records converted per week

Function	Processing time (hours and minutes) for specified numbers of records converted per week						
	5,000	10,000	15,000	20,000	25,000	30,000	35,000
Merge daily verified record tapes[1]	0:03	0:06	0:09	0:12	0:16	0:19	0:22
Generate search code and threaded list and add record to data base	8:20	16:40	25:00	33:20	41:40	50:00	58:20
Select records by user profile	0:08	0:12	0:15	0:19	0:22	0:26	0:30
Duplicate selected records	0:21	0:30	0:39	0:48	0:57	1:06	1:15
Total	8:52	17:28	26:03	34:39	43:15	51:51	60:27

1. In an operating situation merging would probably be a daily operation. Since the merge time depends on file size, however, it is not feasible to calculate the time on this basis. The weekly figures in this table provide an indication of the time that might be required.

Appendix I

STAFF COMPLEMENTS AND UNIT COSTS

Table I.1 presents a detailed analysis of the staff complements for each conversion function for all 20 technical alternatives considered in is study. Only editing and input are true variables. Project direction and quality control are constant for all conversion methods and when catalog comparison applies, the same size staff is required.

Table I.2 gives man/machine costs for each function for the 20 technical alternatives. Here the variations are more evident, ranging from a low of $1.18 (E2) to a high of $2.09 (J4). It is more accurate, however, to compare the low and high figures for conversion without catalog comparison ($1.18 and $1.77) and those for conversion with catalog comparison ($1.51 and $2.09).

Table I.1--Staff complements for each conversion function, by technical alternative

Function	Technical alternative																			
	A2	A3	B2	B3	C2	C3	D2	D3	E2	E3	F2	F3	G2	G3	H1	H4	I1	I4	J1	J4
Project direction	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
Editing	40	43	40	43	40	43	40	43	32	35	32	35	32	35	53	55	53	55	53	55
Input	11	12	21	22	21	22	21	22	20	21	20	21	20	21	25	26	25	26	25	26
Catalog comparison	-	18	-	18	-	18	-	18	-	18	-	18	-	18	-	18	-	18	-	18
Quality control	15	15	15	15	15	15	15	15	15	15	15	15	15	15	15	15	15	15	15	15
Total	70	92	80	102	80	102	80	102	71	93	71	93	71	93	97	118	97	118	97	118

Table I.2--Man/machine unit costs for each function, by technical alternative

Function	Cost per record for each alternative																			
	A2	A3	B2	B3	C2	C3	D2	D3	E2	E3	F2	F3	G2	G3	H1	H4	I1	I4	J1	J4
Project direction	$.090	$.090	$.090	$.090	$.090	$.090	$.090	$.090	$.090	$.090	$.090	$.090	$.090	$.090	$.090	$.090	$.090	$.090	$.090	$.090
Selection																				
Dividing record set	.007	.007	.007	.007	.007	.007	.007	.007	.007	.007	.007	.007	.007	.007	.007	.007	.007	.007	.007	.007
Remerging record set	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003
Preparation																				
Microfilming cards	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002
Making hard copy	.010	.010	.010	.010	.010	.010	.010	.010	.010	.010	.010	.010	.010	.010	.010	.010	.010	.010	.010	.010
Input																				
Keying[1]	.127	.138	.241	.252	.241	.252	.241	.252	.230	.241	.230	.241	.230	.241	.285	.300	.285	.300	.285	.300
Cost of input device[1]	.171[2]	.171[2]	.041	.041	.088	.088	.184	.184	.051	.051	.089	.089	.233	.233	.063	.063	.118	.118	.283	.283
Editing[1]	.597	.640	.597	.640	.597	.640	.597	.640	.478	.521	.478	.521	.478	.521	.788	.816	.788	.816	.788	.816
Format recognition	.047	.047	.047	.047	.047	.047	.047	.047	[3]	[3]	[3]	[3]	[3]	[3]	-	-	-	-	-	-
Output																				
Sorting	-	.009	-	.009	-	.009	-	.009	-	.009	-	.009	-	.009	-	.009	-	.009	-	.009
Printing	.029	.032	.029	.032	.029	.032	.029	.032	.029	.032	.029	.032	.029	.032	.029	.032	.029	.032	.029	.032
Catalog comparison	-	.265	-	.265	-	.265	-	.265	-	.265	-	.265	-	.265	-	.265	-	.265	-	.265
Quality control	.275	.275	.275	.275	.275	.275	.275	.275	.275	.275	.275	.275	.275	.275	.275	.275	.275	.275	.275	.275
Total (rounded)	$1.36	$1.69	$1.34	$1.67	$1.39	$1.72	$1.49	$1.82	$1.18	$1.51	$1.21	$1.54	$1.36	$1.69	$1.55	$1.87	$1.61	$1.93	$1.77	$2.09

---

[1] The unit cost includes any additional work generated by corrections from proofing, catalog comparison, or quality control.

[2] The unit cost for direct-read OCR (.167) has been increased by .004, 10 percent of the machine cost for unedited tape inscriber records (.041) because an estimated 10 percent of the records would be rejected by OCR and thus would have to be input by keyboarding.

[3] The unit cost of format recognition in E2, E3, F2, F3, G2, and G3 is too small (less than .001) to be included in this table.

INDEX

Abel (Richard) & Co., 129
Acceptance sampling, 84
Automation in American libraries, 111-123

Binary search, 204-206, 210
Books for College Libraries, 27, 129
Books in Print, 27, 129
Bucket approach to searching, 204, 210

Catalog comparison: cost, 81, 94, 98, 226; description, 80-83; editing and, 77, 82; justification, 33, 151; printing cost and, 66; recycling and, 217-220; staff, 88-91, 96, 225; technical alternatives using, 11, 46-48
Centralization of conversion, 4, 10, 109, 133
Complexity of catalog records, 55, 78
Content designators, 36, 40-44, 55, 82, 163-182
Computer processing time, 211-216, 220-223; see also Format recognition; Printing; Sorting
Consultants, list of, 134
Conversion of catalog records: benefits, 4, 13, 103; centralization of, 4, 10, 83, 127, 133; consultants' opinions about, 125-127; cost, 97-101, 133, 167; flexible approach to, 55; need for, 1, 13, 125; other libraries' requirements for, 113-116, 118-120, 122, 130; problems, 1-4, 116-118, 126
Conversion priorities, 10, 26, 29-32; consultants' opinions about, 27, 127, 130; costs to implement, 97, 99, 100; exclusions from, 21; other libraries' opinions about, 26, 113, 116, 122
Conversion strategy, 26-29, 127-130
Converter for magnetic tape inscriber, 50, 58
Cost per record; see Unit costs

Disk, 60, 69, 186-191, 201, 206
Distribution service for retrospective records, 3, 121, 132, 165, 184; computer processing time for, 213, 216, 222, 223; cost, 104, 133; LC Card Division and, 28, 101, 104, 132; software, 198, 205-211
Duplication in library collections, 20, 25, 28, 106-109

Editing: catalog comparison and, 77, 82; computer processing time and, 213, 221, 222; consultants' opinions about, 127, 131; cost, 94, 97, 226; definition, 40-42, 75; error rate and, 217-220; examples of, 180-182; format recognition and, 41, 63, 131, 166, 169-171, 176-178, 193, 213, 221; input equipment cost and, 55-58, 61, 97-98; input production rate and, 58, 61; other libraries' experience with, 115-117; software requirements and, 193; staff, 75-77, 87, 90, 225; technical alternatives and, 45-49
Errors, 50, 53, 76, 79, 83-85, 170, 216-220

Filing, 17, 126, 167
File maintenance, 211, 213, 221
File organization, 3, 18, 117, 126, 199-210
Format recognition: algorithms for, 171-178; cost, 64, 98, 226; definition, 42; editing and, 41, 63, 76, 131, 166, 169-171, 176-178, 193, 213, 221; OCR and, 76; processing time for, 63, 213, 221; software, 104, 193, 211; recycling and, 216-220; technical alternatives and, 11, 46-48
Function codes, 170
Funding for conversion, 3, 8, 12, 102-105, 115, 118, 133

Hardware: basic configuration, 185-188; cost, 44, 60, 64, 68-73, 186-191; see also Input devices; Storage
Holdings information, 12, 34-38, 116, 120, 126, 162

Input: cost, 94, 98, 99, 226; description, 77-79; keying rate and, 52, 61, 78; staff, 88, 90, 91, 225
Input devices: cost, 55-63, 226; evaluation of, 49,55, 116; technical alternatives and, 45
Institute of Library Research, 131

Keying; see Input

Language as a factor in conversion, 8, 11, 18, 28, 30-32, 79, 82, 128, 194, 197
Levels of machine-readable records, 16, 36, 118, 163-168; consultants' opinions about, 130-132; definition, 43, 164
Libraries represented in survey, list of, 123
Library of Congress: conversion needs of, 29, 33; funding of conversion effort by, 102-104; policy of changes in catalog records, 80, 147, 156-161; space problems of, 95
LC Card Division, 28, 32, 104, 133, 213
LC Card Division mechanization project, 12, 101, 127
LC Card Division record set: description, 23, 27, 48, 74, 136; OCR and, 52; Official Catalog and, 80-82, 141-152; use of conversion, 11, 23
LC catalog records: bibliographies as a source of, 27, 127; changes in, 22-25, 80-82, 84, 141-162; complexity, 55, 78; consultants' opinions about, 127; format recognition and, 170-178; number, 23, 136-140; OCR and, 52, 59; quality, 2, 21; use by other libraries, 119, 133
LC Official Catalog: conversion of, 26, 47-49; description, 21, 23; master data base and, 11, 22, 32, 128; record set and, 80-82, 141-152
LC shelflist, 25-26, 128, 144, 162
Links, 202-210
List entry, 202-210

Machine-readable records: complexity, 55, 78; content designators for, 36, 40-44, 55, 82, 163-182; format recognition and, 169-179; length, 55-56, 59, 68, 187; levels of, 16, 36, 43, 118, 130-132, 163-168; other libraries' production of, 116; quality, 80, 83-85, 133; standardization, 2, 4, 8, 10, 18, 21, 36, 109, 121, 168
Magnetic tape inscriber: cost, 58, 61, 97; use, 11, 45, 50, 56, 78, 90
Manpower production rates, 40, 76, 94; complexity and, 56, 78, 82; effective working day and, 56, 86
Man-year, 86
MARC Distribution Service: consultants' opinions about, 126; coverage, 1, 10, 30, 102; experience in, 65, 78, 79, 83, 95, 141, 144;

staff, 30, 76, 95; use, 37, 119
MARC II format, 2, 16, 43, 130-132, 163-168; see also Content designators
Master data base, 20-26, 128
Merging; see Sorting
Microfilming, 53, 92-94, 96, 98, 226
Milestone table, 206
MT/ST; see Magnetic tape inscriber
Multiprogramming, 66, 186, 212

National data store: characteristics, 10, 21, 22, 34-38; cost, 104; national union catalog and, 5, 12, 19, 126, 132
National Serials Data Program, 22, 127
National Union Catalog, 21, 34-37, 107; master data base and, 24, 132; reports to, 20, 108, 110

OCR, direct-read: cost, 57, 61, 97, 98; description, 52-54; format recognition and, 76; input staff required, 90; processing rate, 58; reject rate of, 54, 98; software, 73, 100, 104; technical alternative for, 45
OCR scanner, 45, 51, 58, 61, 226
On-demand service, 28, 132, 185, 205-210, 211, 213, 222
On-line typewriter, 45, 51, 59, 62, 226

Permanent index, 201-210
Printing, 46-48, 211; computer processing time for, 65-68, 186, 213, 221; cost, 65-68, 98, 226
Proofing; see Editing

Quality control, 11; cost, 94, 98, 226; description, 83-85; staff, 89-91, 225

RECON study, assumptions of, 7
Record set; see LC Card Division record set
Recycling, 171, 212, 216-220
Retrospective catalog records, definition of, 30

Search code, 132, 199-210
Searching, 13, 18, 29, 213, 222, 223
Selection from data base, 74, 92-94, 96, 98, 226
Serials, 21, 113, 114, 127
Site preparation, 73
Software, 188-210; cost, 5, 44, 73, 100, 211; development time for, 211; funds for, 12, 104; other libraries' experience with, 117
Sorting: catalog comparison and, 81; computer processing time for, 65, 214-216, 221, 223; cost, 64, 98, 226; technical alternatives and, 46-48
Staffing: extent of, 40, 85-91, 95, 105, 224; cost, 92-94; other libraries' experience in, 116-118, 121
Storage, 68-72, 186-191; cost, 44, 60, 68-70, 100; searching methods and, 199-210; sorting and, 214-216
System capacity, 70-72
System design, 3, 7, 12, 100, 104, 115-117, 121

Technical alternatives, 39, 44-49, 62; cost, 93, 98, 99, 226; staff, 89-91, 225
Threaded list, 201, 202, 206-210
Two-up printing, 67, 221

Unit costs: derivation, 55-57, 93; machine, 55-62, 64-68, 97-99, 226; manpower, 75, 92-94, 97-99, 226; other libraries' estimates of, 115
Updating, 3, 24, 118, 141, 151; rate of, 34, 67, 80, 144, 217
User needs, 1, 10, 14, 20, 27-29, 31, 129, 184, 198
User profile, 198, 223

Uses of machine-readable records, 7, 13-19, 114, 119, 121

Verification of machine-readable records; see Quality control

U.S. GOVERNMENT PRINTING OFFICE : 1969 O—351-845